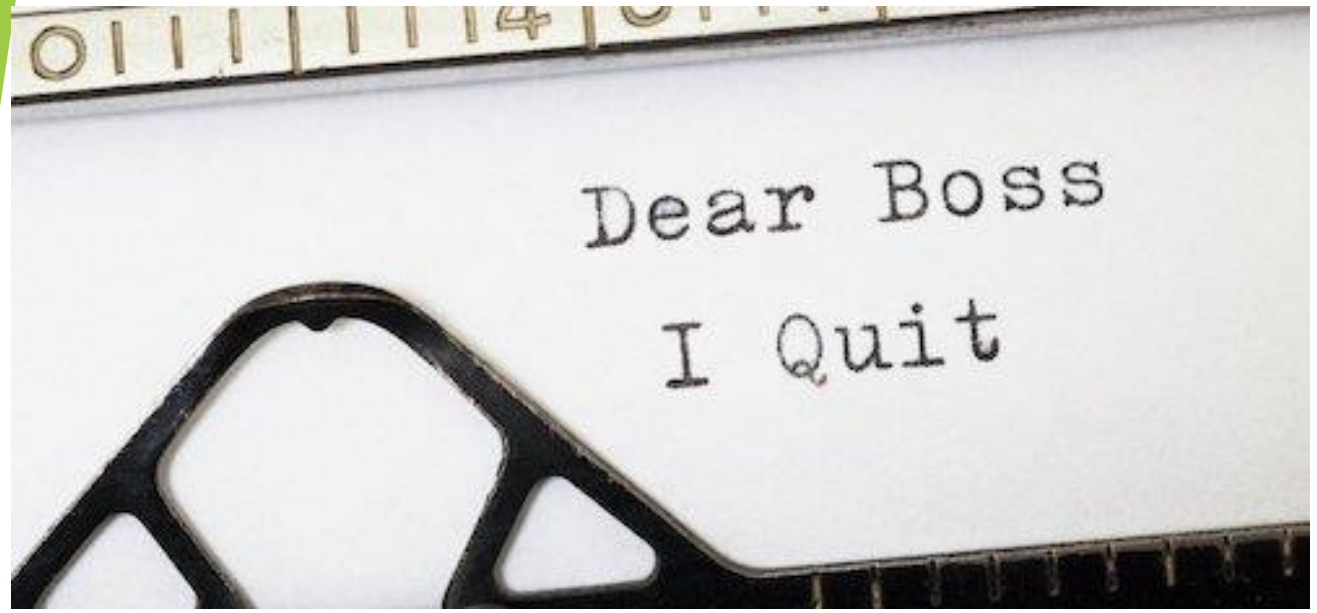
APPLYING MACHINE LEARNING TECHNIQUES TO UNDERSTAND THE "PROBLEM OF EMPLOYEE ATTRITION"



Submitted by:

MD JAVED

Executive Summary

In the second half of 2020, news regarding the “Great Resignation” hit almost all types of media platforms. A survey done by Microsoft reveal that 46% of the workers are planning to leave or switch their job. This is nothing but a type of feedback that organizations have been receiving ever since the covid-19 virus outbreak led to a global lockdown and forced many organizations to either shut their work or run the operations remotely.

This mini-project is an attempt to understand the importance of various factors that drives employees’ decision of voluntarily resigning from their position. We will be using machine learning algorithm to understand the importance of various factors of resignations and try to predict the exit decision of any employee provided the data is available.

Contents

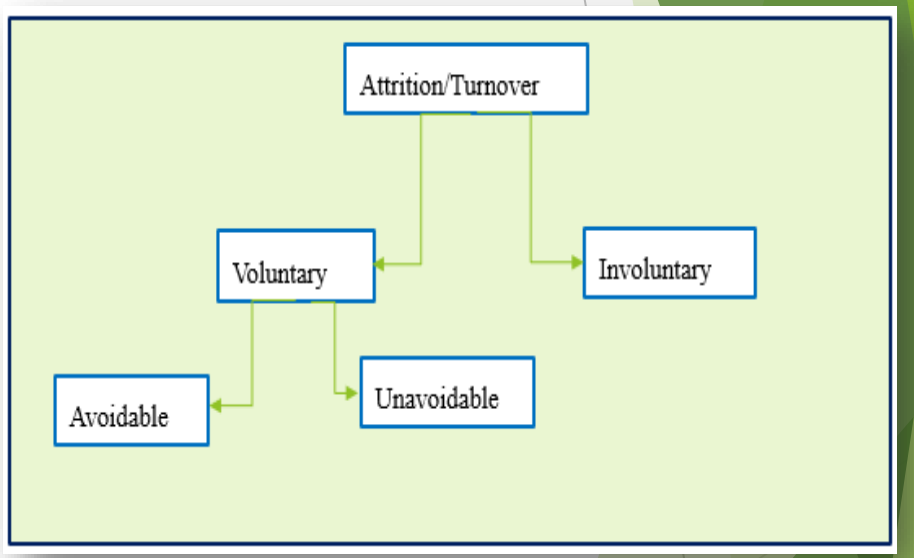
Introduction

Attrition or separation of an employee is the endpoint of every HRM (Human Resource Management) process. Attrition or employee turnover is a situation when an employee either voluntarily or involuntarily leaves the organization. The classification can be understood as:

Voluntary Turnover – It is the situation when the employee willingly initiates the separation process because of compensation issues, workplace culture, poor work-life balance, delayed promotions or lack of growth opportunities, etc.

Involuntary Turnover - It is the situation when the organization initiates the separation because of retirement, performance issues, misconduct at the workplace, loss in profitability, mergers or acquisitions, etc.

The primary goal of every corporate leadership is to create and maintain a workforce environment that has attributes like stability, productivity, collaborative working & superior workforce because this area is critically important to the overall organizational prosperity.



Business Problem

Voluntary turnover is an indicator of the health of any organization, affects the overall business operation adversely & disrupts the human resource budgeting. It is important to note that knowing the intention to leave is a matter of concern.

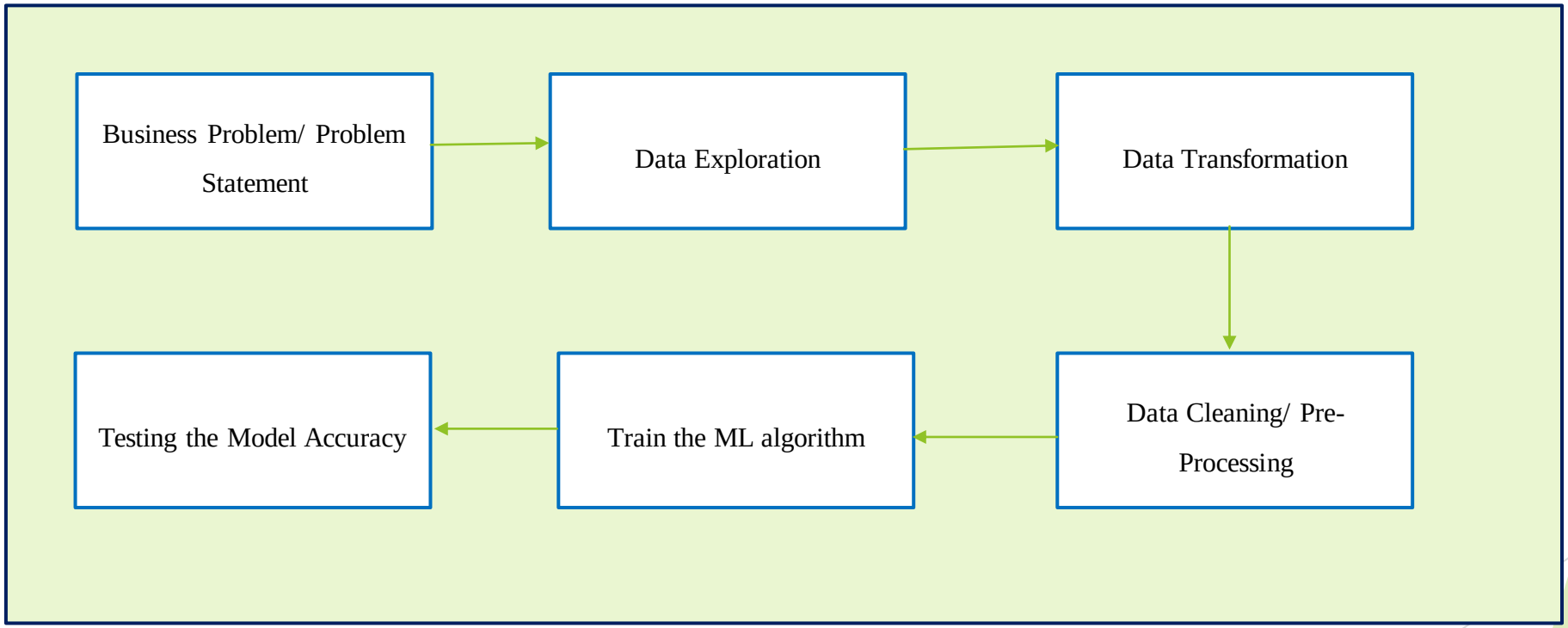
The organization must understand the problem and predict the intent to leave the organization so that company can successfully strategize the retention of good employees.

Objectives:

- To develop data-driven machine learning regression models to predict employees' intention to exit.
- Compare the performance of various models & then select the best performing model.
- We will also look at the factors that play a critical role in formulating the intention to exit.

Methodology

The dataset available in the public domain is used to train the Supervised machine learning algorithm to create a predictive model.



- Problem Statement - To predict employees' intention to resign

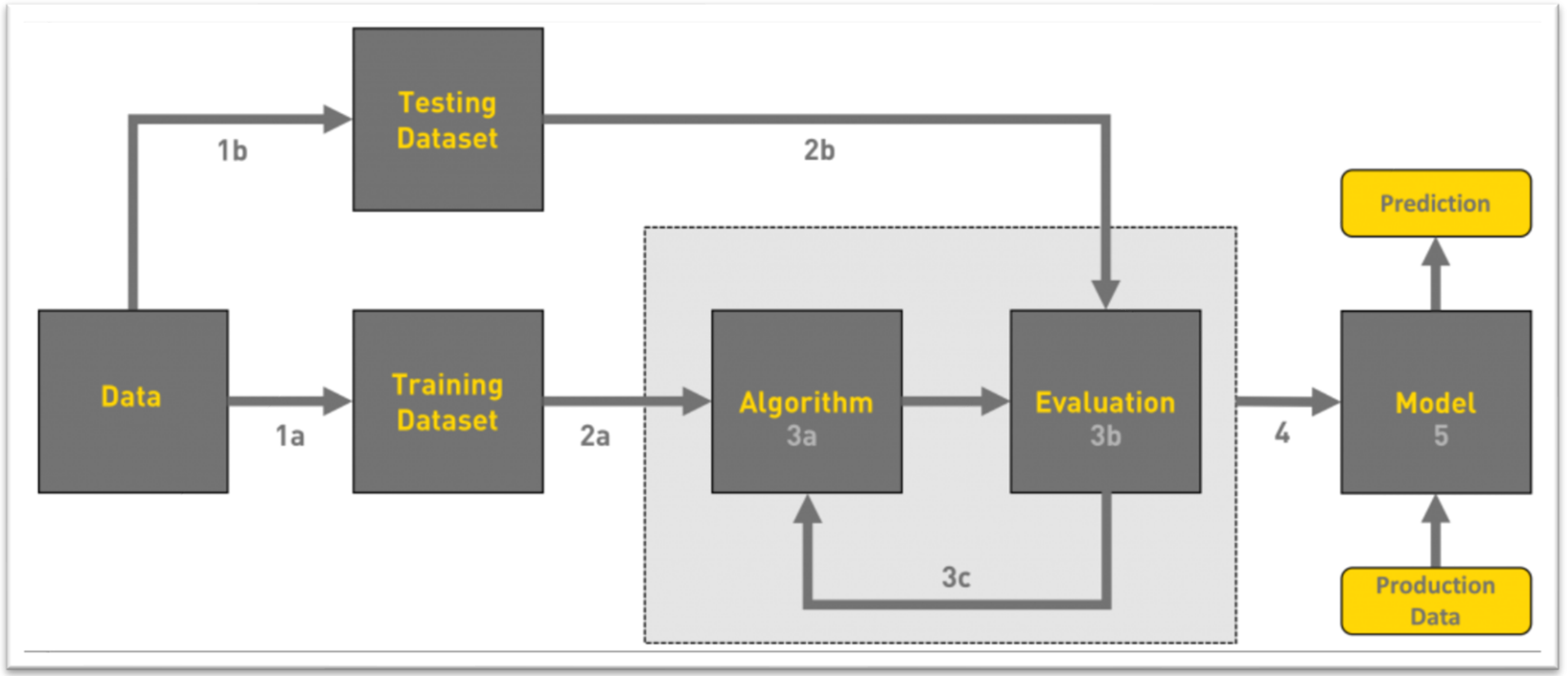
In our project, we will be using the “IBM HR Analytics Employee Attrition & Performance” dataset which is available on the Kaggle website.

ge	Attrition	BusinessT	DailyRate	Departme	DistanceF	Education	Education	Employee	Employee	Environme	Gender	HourlyRat	JobInvolve	JobLevel	JobRole	JobSatisfa	MaritalSta	MonthlyIn	MonthlyR	NumComp	Over18	OverTime	PercentSa	Performar	RelationsI	StandardH	S
41	Yes	Travel_Ra	1102	Sales	1	2	Life Scien	1	1	2	Female	94	3	2	Sales Exec	4	Single	5993	19479	8	Y	Yes	11	3	1	80	
49	No	Travel_Fre	279	Research	8	1	Life Scien	1	2	3	Male	61	2	2	Research	2	Married	5130	24907	1	Y	No	23	4	4	80	
37	Yes	Travel_Ra	1373	Research	2	2	Other	1	4	4	Male	92	2	1	Laboratory	3	Single	2090	2396	6	Y	Yes	15	3	2	80	
33	No	Travel_Fre	1392	Research	3	4	Life Scien	1	5	4	Female	56	3	1	Research	3	Married	2909	23159	1	Y	Yes	11	3	3	80	
27	No	Travel_Ra	591	Research	2	1	Medical	1	7	1	Male	40	3	1	Laboratory	2	Married	3468	16632	9	Y	No	12	3	4	80	
32	No	Travel_Fre	1005	Research	2	2	Life Scien	1	8	4	Male	79	3	1	Laboratory	4	Single	3068	11864	0	Y	No	13	3	3	80	
59	No	Travel_Ra	1324	Research	3	3	Medical	1	10	3	Female	81	4	1	Laboratory	1	Married	2670	9964	4	Y	Yes	20	4	1	80	
30	No	Travel_Ra	1358	Research	24	1	Life Scien	1	11	4	Male	67	3	1	Laboratory	3	Divorced	2693	13335	1	Y	No	22	4	2	80	
38	No	Travel_Fre	216	Research	23	3	Life Scien	1	12	4	Male	44	2	3	Manufact	3	Single	9526	8787	0	Y	No	21	4	2	80	
36	No	Travel_Ra	1299	Research	27	3	Medical	1	13	3	Male	94	3	2	Healthcar	3	Married	5237	16577	6	Y	No	13	3	2	80	
35	No	Travel_Ra	809	Research	16	3	Medical	1	14	1	Male	84	4	1	Laboratory	2	Married	2426	16479	0	Y	No	13	3	3	80	
29	No	Travel_Ra	153	Research	15	2	Life Scien	1	15	4	Female	49	2	2	Laboratory	3	Single	4193	12682	0	Y	Yes	12	3	4	80	
31	No	Travel_Ra	670	Research	26	1	Life Scien	1	16	1	Male	31	3	1	Research	3	Divorced	2911	15170	1	Y	No	17	3	4	80	
34	No	Travel_Ra	1346	Research	19	2	Medical	1	18	2	Male	93	3	1	Laboratory	4	Divorced	2661	8758	0	Y	No	11	3	3	80	
28	Yes	Travel_Ra	103	Research	24	3	Life Scien	1	19	3	Male	50	2	1	Laboratory	3	Single	2028	12947	5	Y	Yes	14	3	2	80	
29	No	Travel_Ra	1389	Research	21	4	Life Scien	1	20	2	Female	51	4	3	Manufact	1	Divorced	9980	10195	1	Y	No	11	3	3	80	
32	No	Travel_Ra	334	Research	5	2	Life Scien	1	21	1	Male	80	4	1	Research	2	Divorced	3298	15053	0	Y	Yes	12	3	4	80	
22	No	Non-Trave	1123	Research	16	2	Medical	1	22	4	Male	96	4	1	Laboratory	4	Divorced	2935	7324	1	Y	Yes	13	3	2	80	
53	No	Travel_Ra	1219	Sales	2	4	Life Scien	1	23	1	Female	78	2	4	Manager	4	Married	15427	22021	2	Y	No	16	3	3	80	
38	No	Travel_Ra	371	Research	2	3	Life Scien	1	24	4	Male	45	3	1	Research	4	Single	3944	4306	5	Y	Yes	11	3	3	80	
24	No	Non-Trave	673	Research	11	2	Other	1	26	1	Female	96	4	2	Manufact	3	Divorced	4011	8232	0	Y	No	18	3	4	80	
36	Yes	Travel_Ra	1218	Sales	9	4	Life Scien	1	27	3	Male	82	2	1	Sales Rep	1	Single	3407	6986	7	Y	No	23	4	2	80	
34	No	Travel_Ra	419	Research	7	4	Life Scien	1	28	1	Female	53	3	3	Research	2	Single	11994	21293	0	Y	No	11	3	3	80	
21	No	Travel_Ra	391	Research	15	2	Life Scien	1	30	3	Male	96	3	1	Research	4	Single	1232	19281	1	Y	No	14	3	4	80	
34	Yes	Travel_Ra	699	Research	6	1	Medical	1	31	2	Male	83	3	1	Research	1	Single	2960	17102	2	Y	No	11	3	3	80	
53	No	Travel_Ra	1282	Research	5	3	Other	1	32	3	Female	58	3	5	Manager	3	Divorced	19094	10735	4	Y	No	11	3	4	80	

Data Overview

#	Column	Non-Null Count	Dtype
0	Age	1470 non-null	int64
1	Attrition	1470 non-null	object
2	BusinessTravel	1470 non-null	object
3	DailyRate	1470 non-null	int64
4	Department	1470 non-null	object
5	DistanceFromHome	1470 non-null	int64
6	Education	1470 non-null	int64
7	EducationField	1470 non-null	object
8	EmployeeCount	1470 non-null	int64
9	EmployeeNumber	1470 non-null	int64
10	EnvironmentSatisfaction	1470 non-null	int64
11	Gender	1470 non-null	object
12	HourlyRate	1470 non-null	int64
13	JobInvolvement	1470 non-null	int64
14	JobLevel	1470 non-null	int64
15	JobRole	1470 non-null	object
16	JobSatisfaction	1470 non-null	int64
17	MaritalStatus	1470 non-null	object
18	MonthlyIncome	1470 non-null	int64
19	MonthlyRate	1470 non-null	int64
20	NumCompaniesWorked	1470 non-null	int64
21	Over18	1470 non-null	object
22	OverTime	1470 non-null	object
23	PercentSalaryHike	1470 non-null	int64
24	PerformanceRating	1470 non-null	int64
25	RelationshipSatisfaction	1470 non-null	int64
26	StandardHours	1470 non-null	int64
27	StockOptionLevel	1470 non-null	int64
28	TotalWorkingYears	1470 non-null	int64
29	TrainingTimesLastYear	1470 non-null	int64
30	WorkLifeBalance	1470 non-null	int64
31	YearsAtCompany	1470 non-null	int64
32	YearsInCurrentRole	1470 non-null	int64
33	YearsSinceLastPromotion	1470 non-null	int64
34	YearsWithCurrManager	1470 non-null	int64

- Data Processing & Model Building Plan



- Data Transformation – We, generally, do not get the data in a predefined format. So, we may require to transform the data from another format to standard CSV (comma-separated values) format. However, we do require to perform this step as our data is already in CSV format.
- Data Pre-processing –

- We have:

  a) Checked for the null values

  b) Removed the unnecessary features from the dataset

  c) Performed binary coding of the categorical variables

  d) Created the training & testing data by splitting the original dataset

  e) Features normalization by using Min-Max scaling method

  f) Applying SMOTE (Synthetic Minority Over-Sampling Technique) to **up sample** the minority attrition data)

- Train the Machine Learning Algorithms:

  a) Decision Tree Algorithm

  b) Random Forest Classification Algorithm

  c) K-Nearest Neighbours Classification Algorithm

  d) Logistic Regression Classification Algorithm

Removed Columns

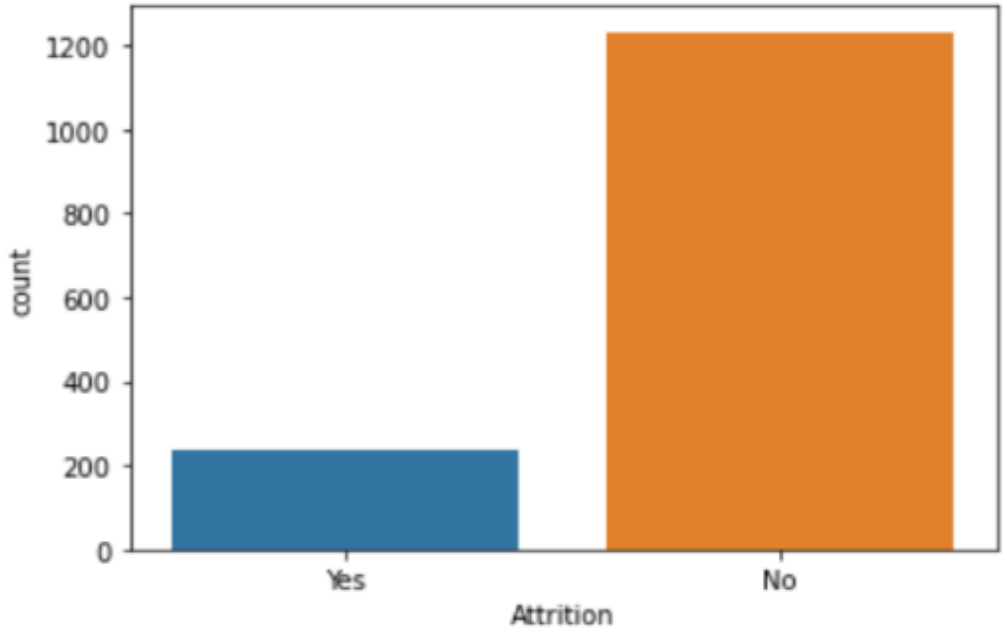
```
#Remove the column EmployeeNumber
df = df.drop('EmployeeNumber', axis = 1) # A number assignment
#Remove the column StandardHours
df = df.drop('StandardHours', axis = 1) #Contains only value 80
#Remove the column EmployeeCount
df = df.drop('EmployeeCount', axis = 1) #Contains only the value 1
#Remove the column EmployeeCount
df = df.drop('Over18', axis = 1) #Contains only the value 'Yes'
```

```
#Remove column Joblevel
df = df.drop('JobLevel', axis = 1) #sharing 95% correlation with 'MonthlyIncome' column
```

Data Analysis - Exploratory Data Analysis

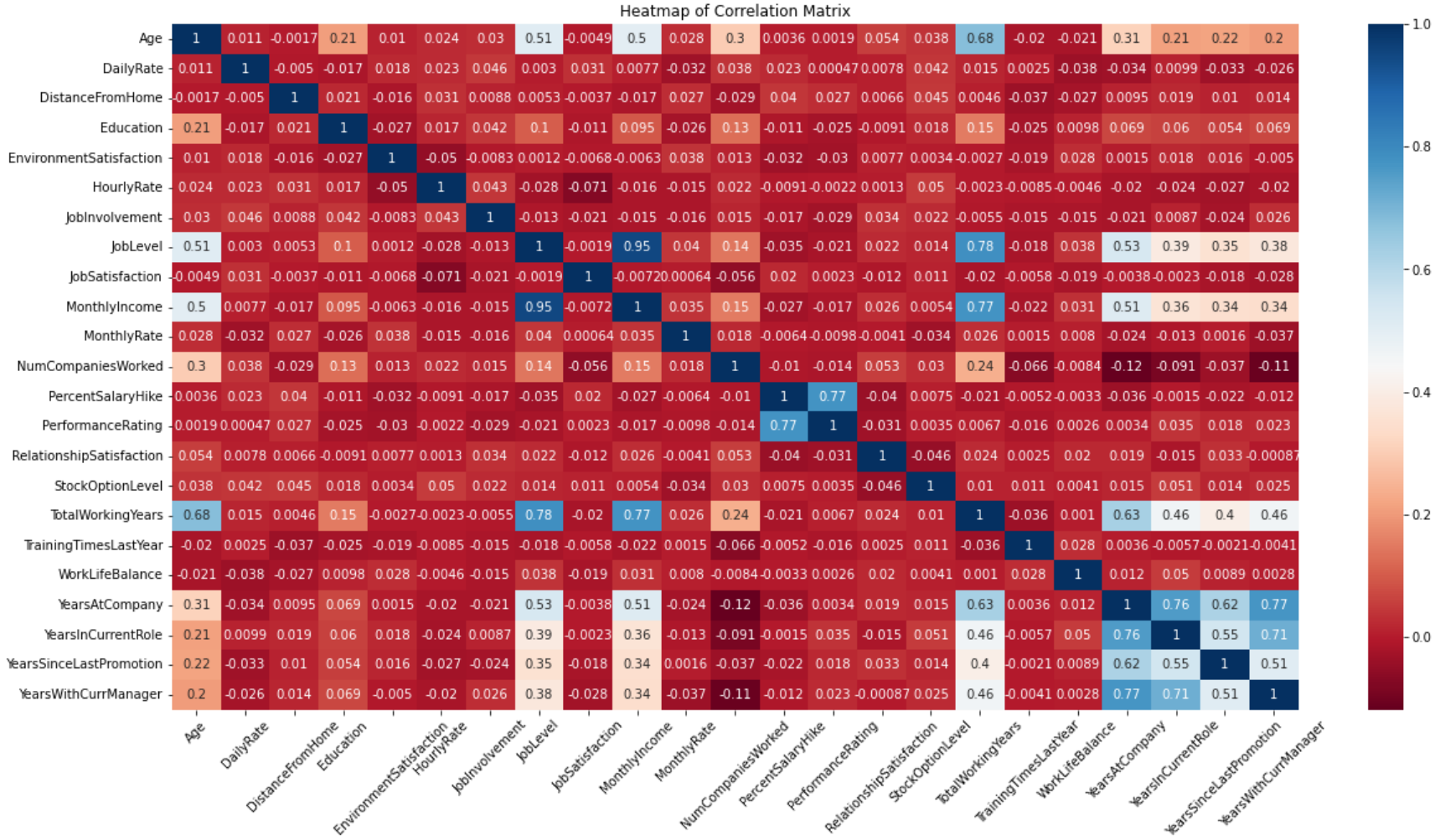
Overall Attrition Count in the Data -

We have an Attrition rate of 16.22% by using the formula: count (Attrition)/total records.



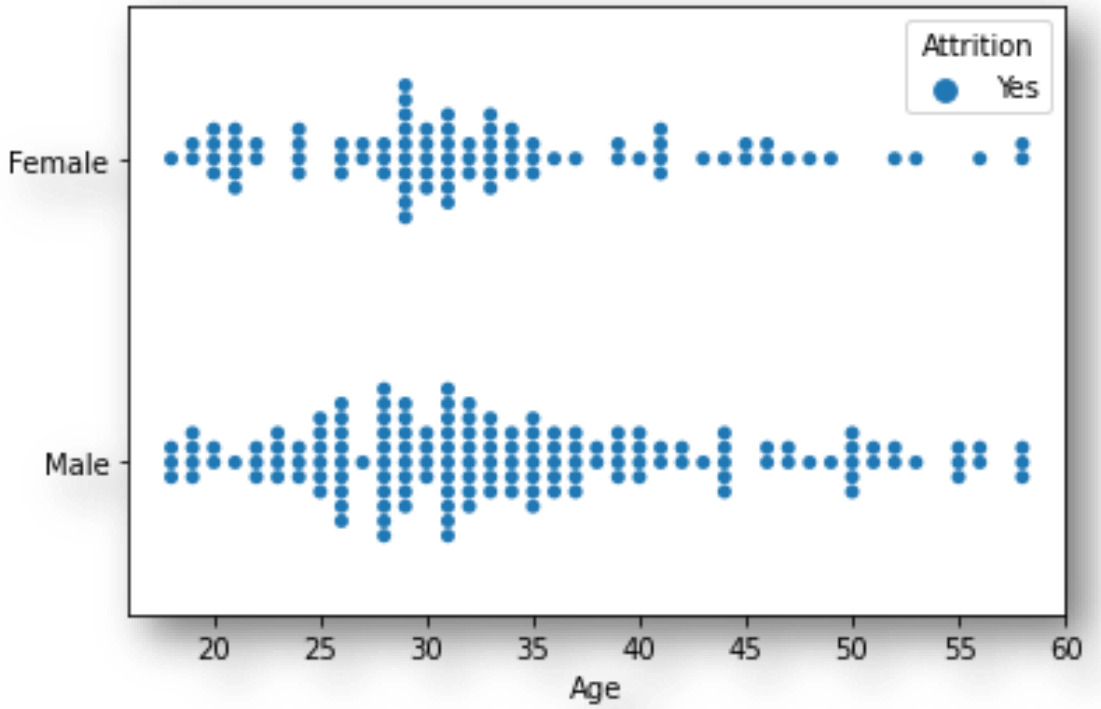
The distribution of attrition show imbalanced distribution in the sample. To get a good result with machine learning technique, the samples are oversampled by using SMOTE (Synthetic Minority Over-Sampling) method.

Overall Data Heatmap:



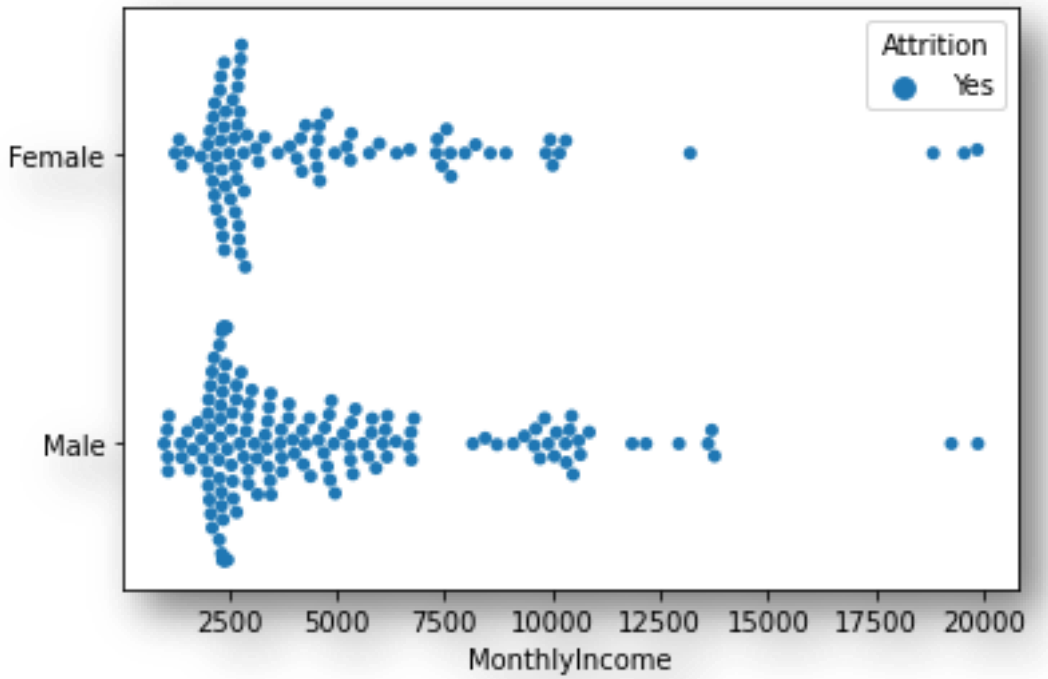
We can observe 95% correlation between “JobLevel” & “MonthlyIncome”. We drop “JobLevel” as income is very important feature. Otherwise, this strong multicollinearity will create problem to identify the important features.

Age and Gender-wise Attrition Spread-



The high Attrition rate is clustered around the age of 30 years that is young employees tend to switch companies more often. There is no significant difference between the distribution of attrition across the gender.

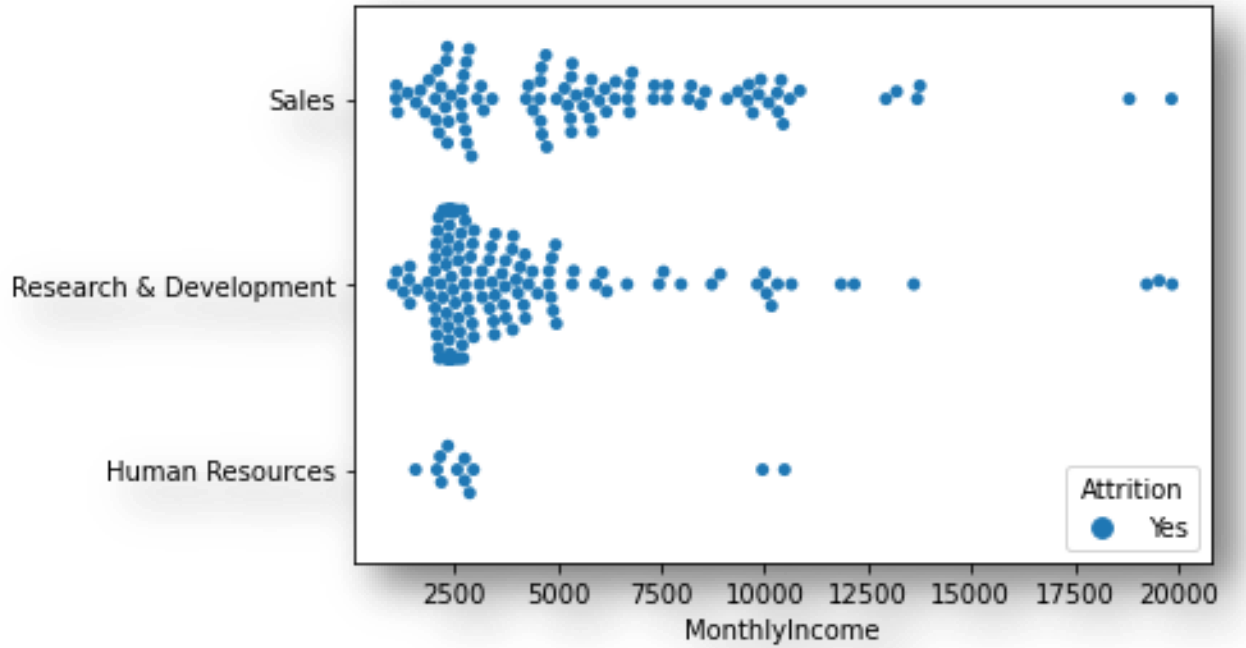
Monthly Income-wise Attrition Spread with Gender effect:



We can observe that:

- More attrition instances are recorded in the group of Male
- In Male employees, the considerable attrition instances till USD 5,000.
- However, Females are more clustered around USD 2,500.
- Hence, Male employee with income less than USD 5,000 is more likely to leave the current organization as compared to their female counterpart.

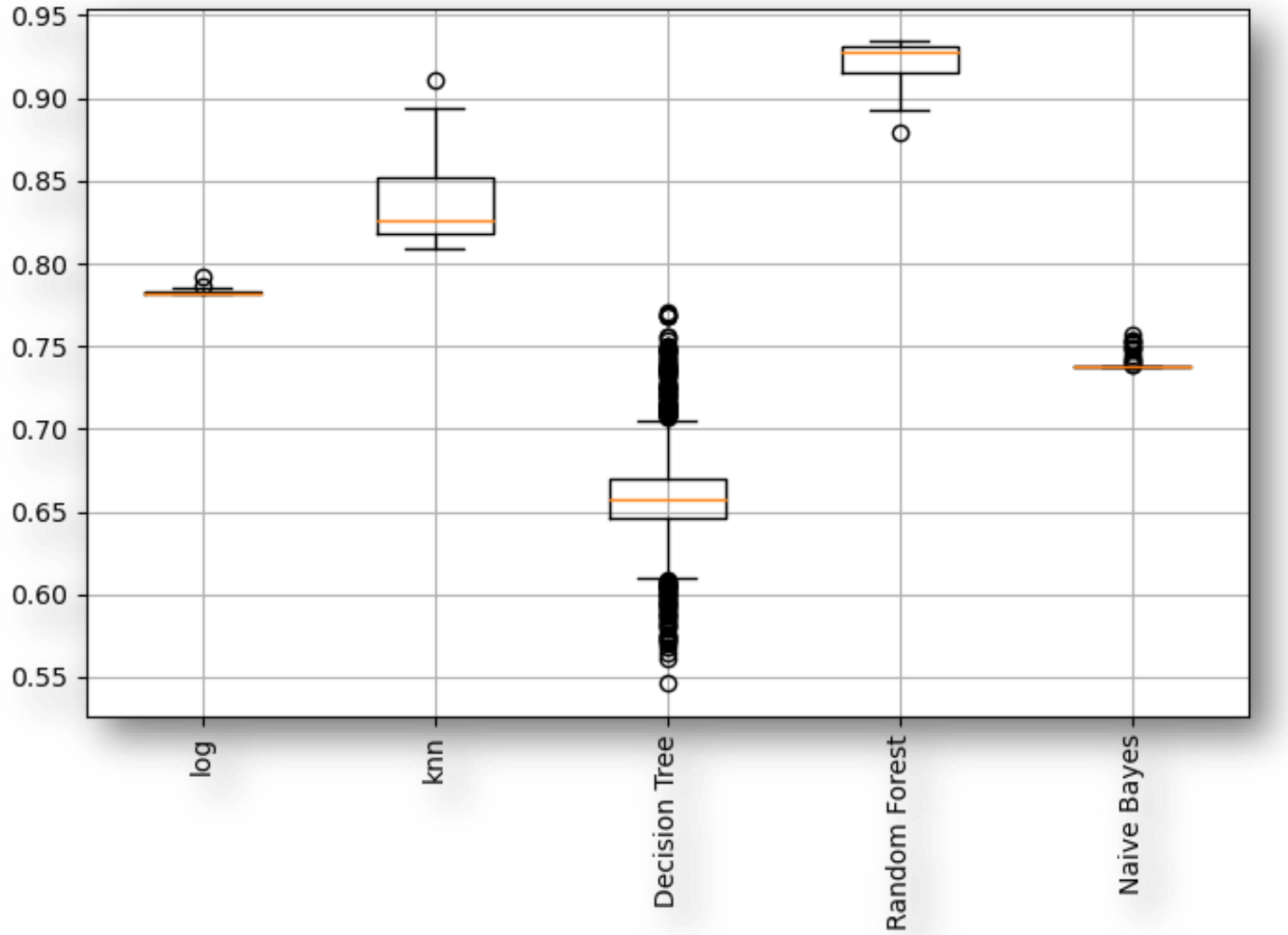
Monthly Income-wise Attrition Spread across various Departments:



We can observe that:

- In R&D department, there's considerable attrition instances till USD 5,000.
- HR has least attrition cases clustered around low income level.
- In Sales, attrition is more in both low & middle income group.
- Our employees from R&D department are getting more lucrative opportunities after learning skills on the job and poses higher flight risk (risk of leaving the organization voluntarily).

Model Results:

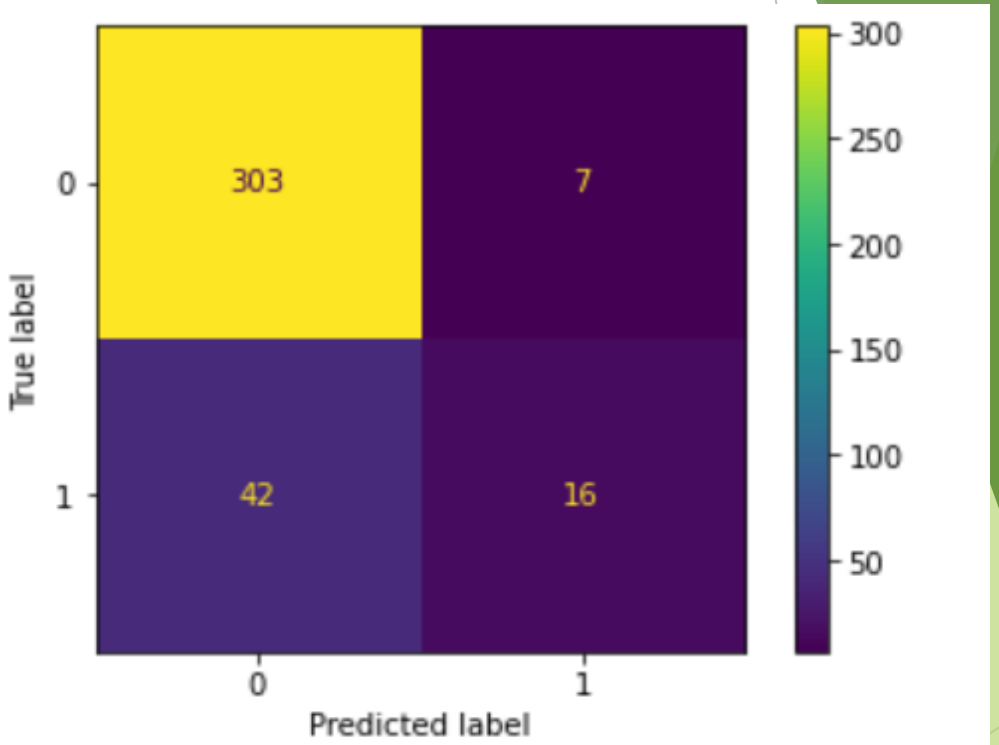


We can clearly see that our trained model based on the ‘Random Forest’ (RCF) algorithm has yielded highest accuracy.

Classification Report & the Confusion Matrix for RCF model:

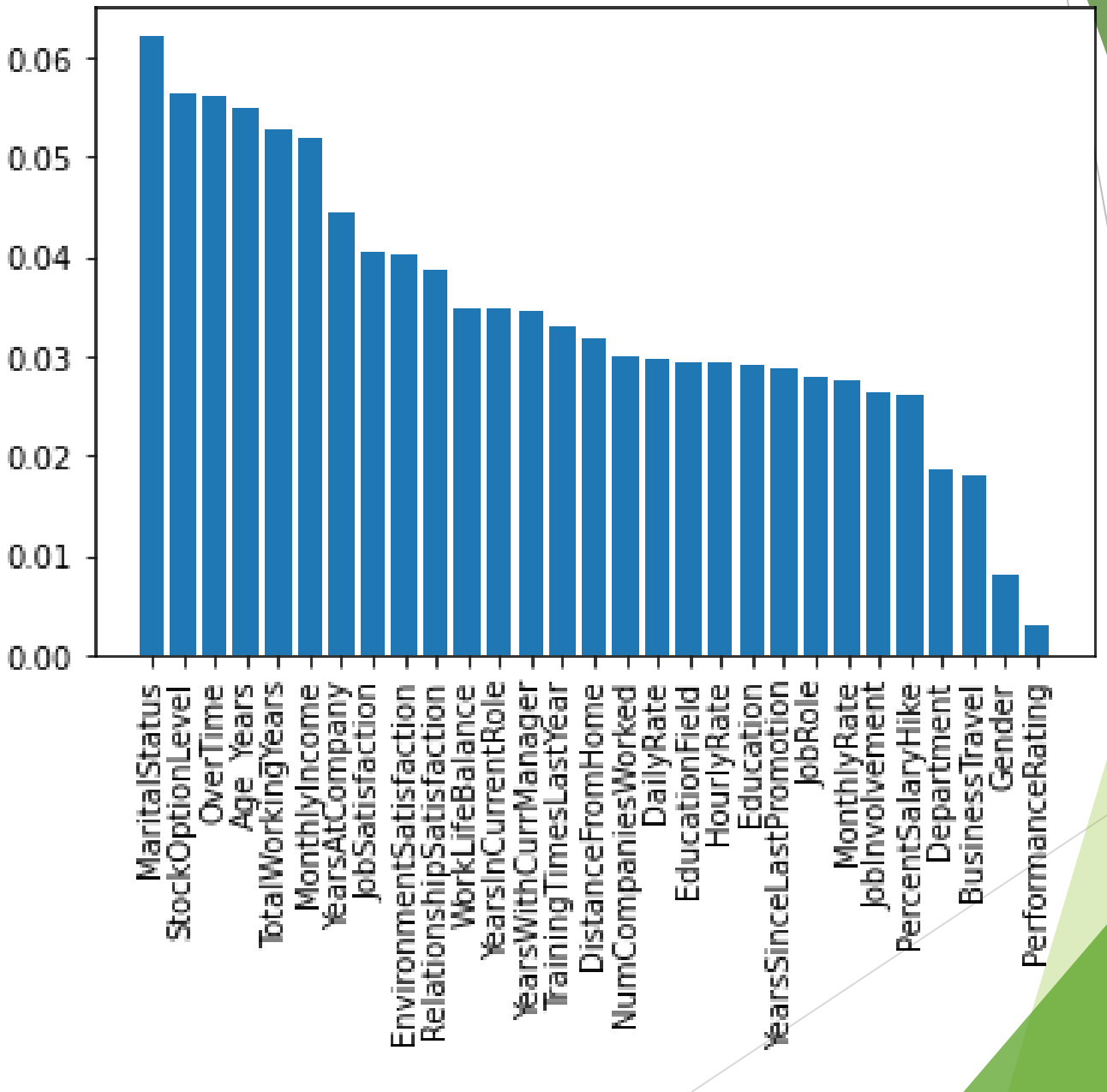
```
print(classification_report(Y_test, Y_RF_pred))
```

	precision	recall	f1-score	support
0	0.88	0.98	0.93	310
1	0.70	0.28	0.40	58
accuracy			0.87	368
macro avg	0.79	0.63	0.66	368
weighted avg	0.85	0.87	0.84	368



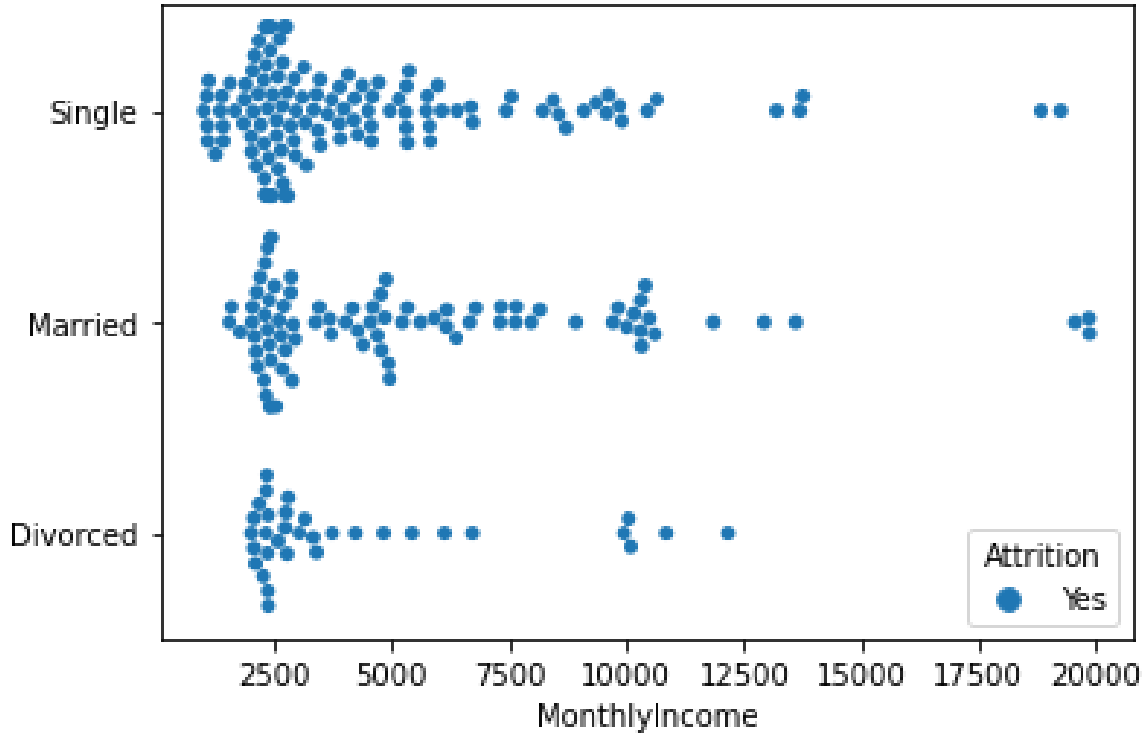
We can clearly observe that our model is identifying the intention to quit (value = 1 in classification report) with 70% precision.

Feature Importance of Random Forest Classifier



Observations:

- 'MaritalStatus' of employee is playing a major role in defining their intention to quit followed by 'StockOption' & 'OverTime'.
- Factors 'StockOption' & 'OverTime' indicating the monetary reason.



we observe that:

- Single employees who are earning less than USD 5,000 are showing high degree of intention to leave.
- Married employees group follow the similar pattern as Single employee group does but the number of incidents and hence the likelihood is less.
- For Divorced employees, the threshold is USD 3,000 that shows they are less likely to leave in comparison to the single employees group once their income crosses this threshold.

Conclusion, Recommendations & Limitations:

We recommend the following strategies to manage the situation:

Talent Acquisition:

Identify:

- Candidates that are more interested in the organization culture
- Candidates who shows high degree of long term commitments
- Candidates who are not single because they show relatively more scope of stability than their Singles counterparts

Compensation & other benefits:

- Additional compensation opportunities can be provided in the form of “OverTime” for new employees at the low level of income & “StockOption” for employees who are tenured but works at lower income level.
- It is important to note that the company should perform the performance evaluation before rolling out such options otherwise, everyone will get such opportunity including the least efficient employees and hence, company loses its money/resources.
- Giving non monetary incentives like “awards & recognition” to star employees at lower earning level to influence their intention to quit/or not to quit

Limitation & further improvements:

- Since, the data is more skewed towards the “non-attrition” category. Therefore, our model has the inherent bias towards predicting no attrition. Oversampling/under sampling techniques can be deployed in such situations but it has its own limitations in terms of reliability.
- For improvement, more data on attrition should be collected. The more refined & rich our input is, better output from the machine learning model we can expect.
- Model can be deployed on server & concerned groups can access the performance through a user friendly UI.
- Separate models for different gender groups or departments can be created to understand the in depth dynamics. We have limited ourselves to the overall analysis because of the data limitation.
- No machine learning model is perfect as it require constant fine tuning. Hence, there should be a proper model performance evaluation mechanism to establish its effectiveness.

Thank You