VIT-AP UNIVERSITY

STATISTICAL PROJECT

ON

IBM HR EMPLOYEE ATTRITION AND PREDICTION

Submitted to : Dr. SANTANU MANDAL

Submitted by :

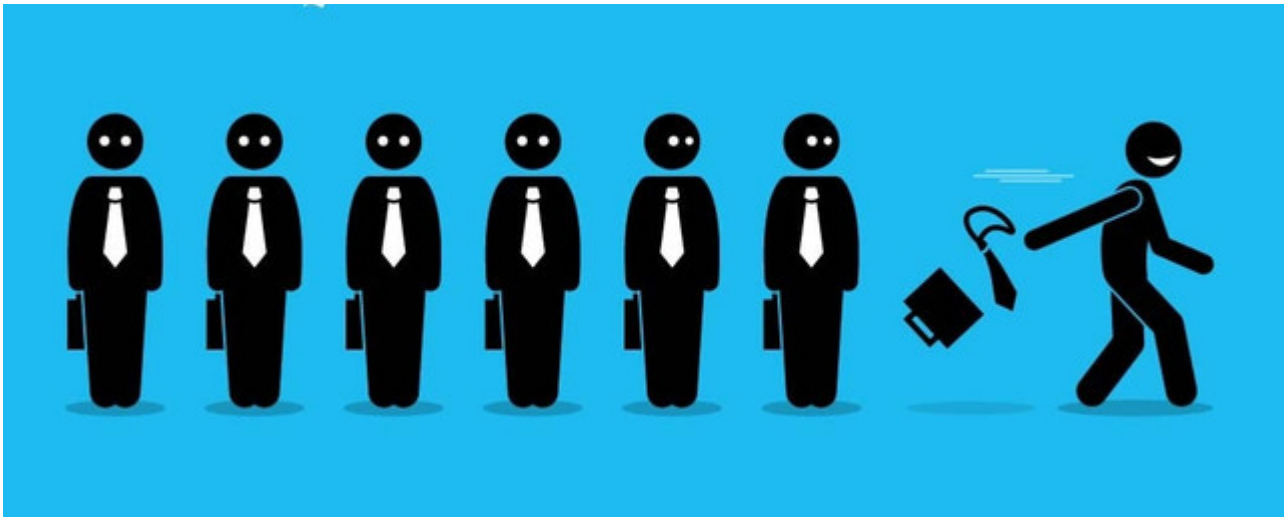
P.VINOD[22MSD7031]

IBM HR EMPLOYEE ATTRITION AND PREDICTION

ABSTRACT:

This report comprises exhaustive research on Kaggle's "IBM HR Analytics Employee Attrition & Performance" data. The study consists of identifying the root cause of the problem of attrition, finding several indicators and features which are leading to such issues and expansive visualisations which adhere to this problem faced by the IBM Human Resource team. The study also entails Random Forest Classifier which point towards some of the leading features which contribute towards attrition. In the end the report concludes with some of the recommendations regarding features which need special attention by the department and practices which can be adopted for better performance. Business is not just doing deals; business is having great products, doing great engineering, and providing tremendous service to customers.

Key words: Attrition, Human Resource, Age, Overtime, Environment



INTRODUCTION:

In the field of Human Resources, HR, when employees decide to quit, this is referred to as employee attrition, and this is the focus of our analysis. In looking further into the causes of attrition. We see that there is overlapping evidence for why employees decide to leave, where in most cases it is due to the following reasons: unsatisfying compensation, unsatisfactory benefits, lack of growth or development opportunities, issues with work-life balance, poor management, poor work conditions, and lack of recognition for work accomplishments or value add in the workplace.

New industry developments and changes in the field may mean that there is no longer a need for certain roles to be filled. This phenomenon occurs most often in declining industries, for instance,

print media, or in fields where new technological developments mean automation can replace manpower. Organisations often choose to scale down the workforce by closing out the employee lifecycle when it reaches its natural end instead of starting anew.

Employee capital is one of the greatest assets an organisation can possess. Companies can spend as much as 70% of total business costs on employees. These costs include salaries, training, recruitment, and skill investments. Furthermore, recruiting and keeping top talent is important to the growth and long-term viability of any company. Often employees hold key characteristics that are instrumental in moving the company forward. Knowing this, when employees decide to quit or leave a company, it can be a serious issue. With each employee, the company loses its direct investment along with all the knowledge and experience that the employee would have inherently provided.

Noting the subtle differences between employee attrition and employee turnover can help leaders strategize more effectively. While turnover requires an investment into recruitment and training, attrition offers an opportunity to trim costs and reimagine organisational operations. Leaders can take the company in a new direction without resorting to layoffs or causing staff discontent.

However, attrition should always be a guided process. It is the job of management to guide the change. Leaders can collect employee input on the process. However, managers should never leave it solely to employees to pick up the slack without any instruction and figure out how to operate with a smaller staff. Attrition should be strategic, and should involve communication.

METHODOLOGY:

(Data source: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.)

The data is provided by IBM on Kaggle. After processing the dataset and cleaning the inconsistencies, the numerical and categorical features used in the attrition prediction model are generated. Various Classification algorithms are used to predict attrition based on set of independent variables like gender, workplace distance, stock level, salary hike etc. used. The predictive models are also used to identify the variables that strongly influence the attrition using variable importance and probabilistic approaches. The models are evaluated using relevant model performance measures to arrive at the most robust models for prediction.

This data set presents an employee survey from IBM, indicating if there is attrition or not. The data set contains approximately 1500 entries. Given the limited size of the data set, the model should only be expected to provide modest improvement in identification of attrition vs a random allocation of probability of attrition. While some level of attrition in a company is inevitable, minimising it and being prepared for the cases that cannot be helped will significantly help improve the operations of most businesses.

As a future development, with a sufficiently large data set, it would be used to run a segmentation on employees, to develop certain “at risk” categories of employees. This could generate new insights for the business on what drives attrition, insights that cannot be generated by merely informational interviews with employees. IBM has gathered information on employee satisfaction, income, seniority and some demographics. It includes the data of 1470 employees. To use a matrix structure, we changed the model to reflect the following data.

OBJECTIVES:

· To know the satisfactory level of employees towards their job and working conditions

· To identify the factors which make employees dissatisfied about company’s policy and norms.

· To find the areas where companies is lagging behind

· To know the reasons why attrition occurs in companies.

· To find ways to reduce attrition in companies.

Dataset Description:

Table 1 – Numerical features used in the user attrition analysis model

Feature Name	Feature Description	Min value	Max Value	Std deviation
Age	Age of Employee	18	60	9.3
Daily Rate	It is the billing cost for an individual's services for a single day	102	1499	403.50
DistanceFrom Home	It is the distance between the company and home of the employee	1	29	8.9
Employee Count	Count of the Employee	1	1	0.0
Education	Education qualification of the employee	1	5	1.02
EmployeeNumber	It is a unique number that has been assigned to each current and former employee	1	2068	602.2
Employee satisfaction	It is all about an individual's feelings about the work environment and organisational culture.	1	4	1.09

HourlyRate	The amount of money that is paid to an employee for every hour worked	30	100	20.32
JobInvolvement	Job satisfaction happens when an employee feels he or she is having job stability.	1	4	1.10
MonthlyIncome	Gross monthly income is the amount of income an employee earns in one month.	1009	19999	4707.95
MonthlyRate	If a monthly rate is set, employees should be paid in exchange for normal hours of work of a full-time worker.	2094	2699	7117.78
NumCompanies Worked	Number of other companies the employee previously worked for.	0	9	2.49
PercentSalaryHike	The amount a salary is increased of an employee in percentage	11	25	3.65
Performance Rating	Rating means gauging and comparing the performance.	3	4	0.36
Relationship Satisfaction	It is the rate of satisfaction between employer–employee relationship.	1	4	1.08
StandardHours	Standard Hour the Employee is Working	80	80	0.00
StockOptionLevel	Employee Stock Option	0	3	0.85
TotalWorkingYear	Total number of years employee worked	0	40	7.78
TrainingTimesLastYear	Number of months the employee is trained by the company.	0	6	1.28
Work Life Balance	Work-life balance refers to the level of prioritisation between personal and professional activities .	1	4	0.70
YearsAtCompany	Total Number of Years at the Company	0	40	6.12
YearsInCurrentRole	Number of years employee worked in current role	0	18	3.62
Years Since Last Promotion	Number of years of an employee since last promotion	0	15	3.22

Years with Current Manager	Number of years employee worked with current manager	0	17	3.56

Table1: It shows the Numerical features along with the Statistical Parameters.

Table2:Categorical Features Used in the user Attrition Analysis Model

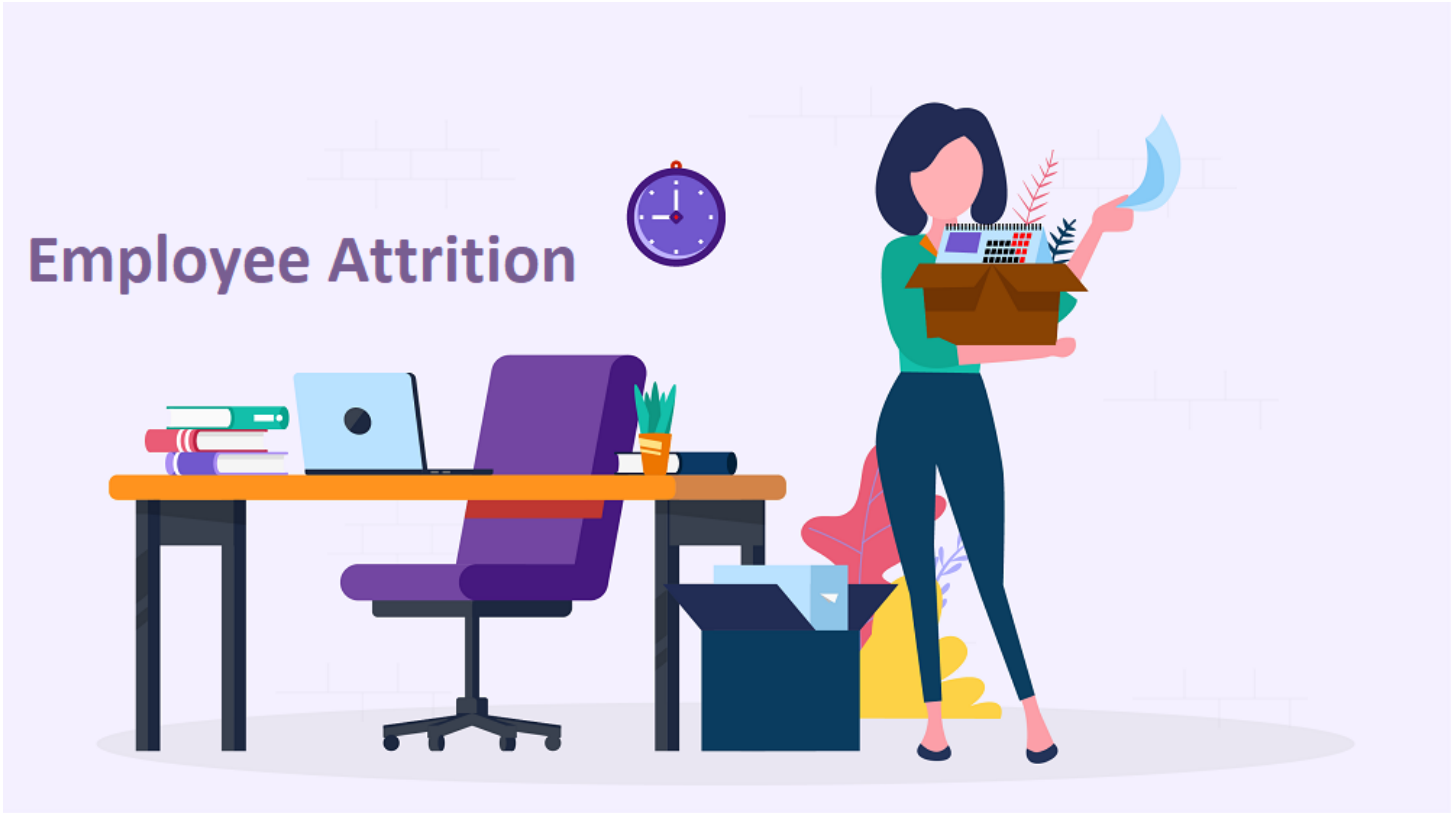
Feature Name	Feature Description	Number of Categorical Variables
Attrition	Attrition in business describes a gradual but deliberate reduction in staff numbers that occurs as employees retire or resign, [NOTE: Target Variable] (0=no, 1=yes)	2
Business Travel	Business travel is travel undertaken for work or business purposes, as opposed to other types of travel (1=NoTravel, 2=Travel Frequently, 3=Travel Rarely)	3
Department	Consists three departments that contribute to the company's overall mission. (1=HR, 2=R&D, 3=Sales)	3
EducationField	Education field of the employees(1=HR, 2=LifeSciences, 3=Marketing, 4=Medical Sciences, 5=others,6= Technical)	6
Job Role	These refer to the specific activities or work that the employees will perform. (1=HC Rep, 2=HR,3=LabTechnician,4=manager, 5= Managing Director, 6= Research Director, 7= Research Scientist, 8=sales Executive, 9= Sales Representative)	9

Exploratory data analysis:

The purpose of exploratory data analysis is :

To understand the data in terms of attrition information across various independent variables.

Get insights on various features



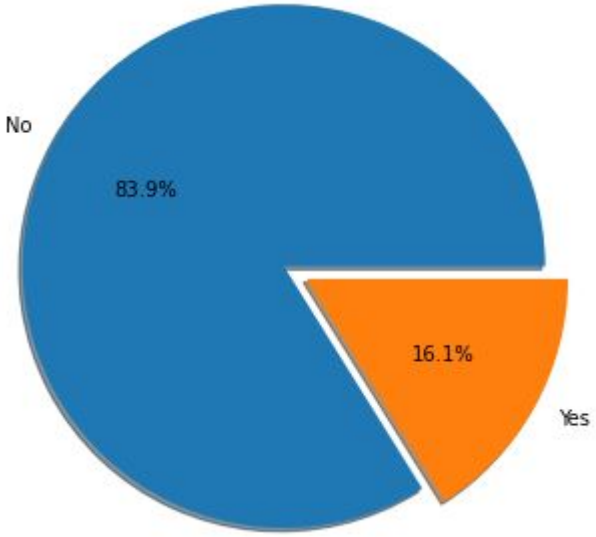
.

TOOLS AND ANALOGY:

Python is the key tool used in our project. Data analysts and experts can use Python as a powerful tool to do complex statistical calculations, produce data visualisations, develop machine learning algorithms, manage, and analyse data.

We use many python libraries like pandas, numpy ,sklearn,seaborn,matplotlib,plotly….e.t.c;

1.Overall Attrition

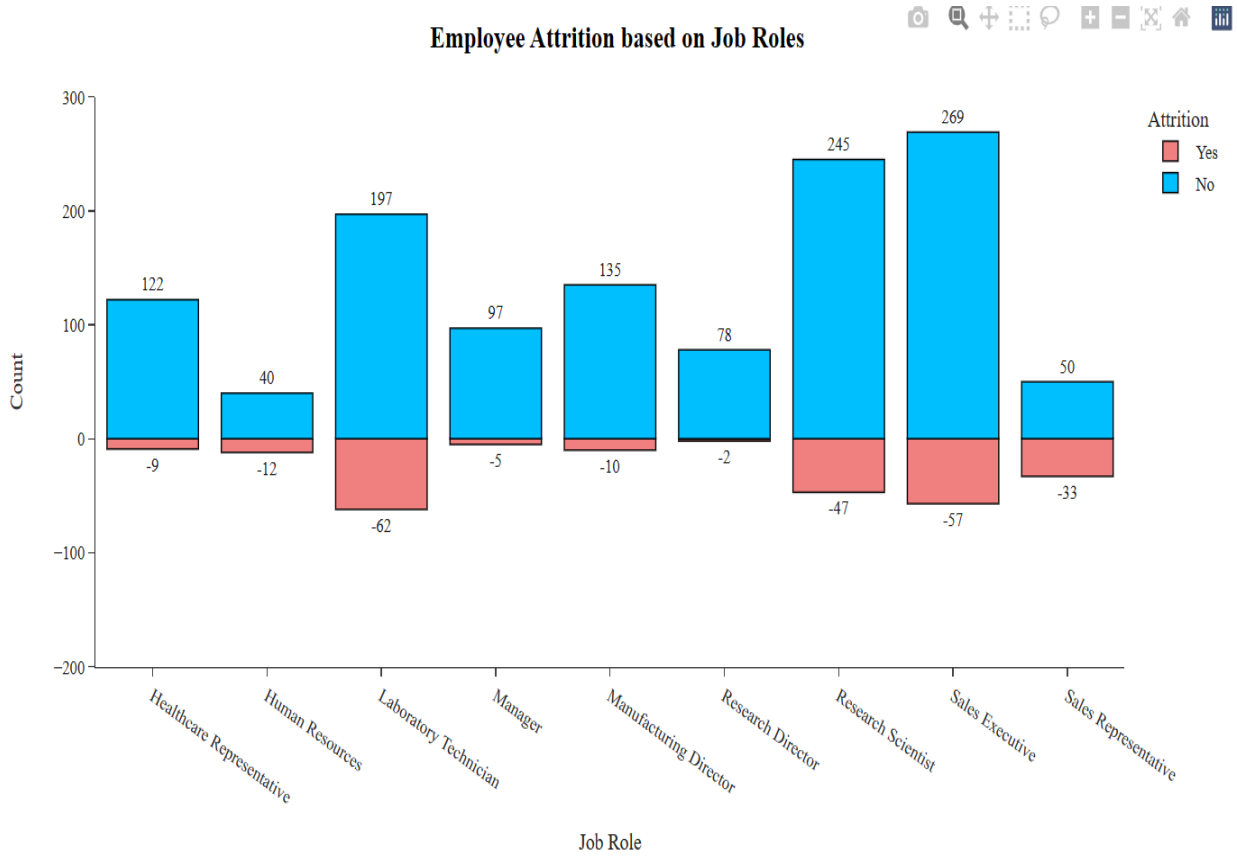


The attrition rate is = 237/1470 = 16.1 %.

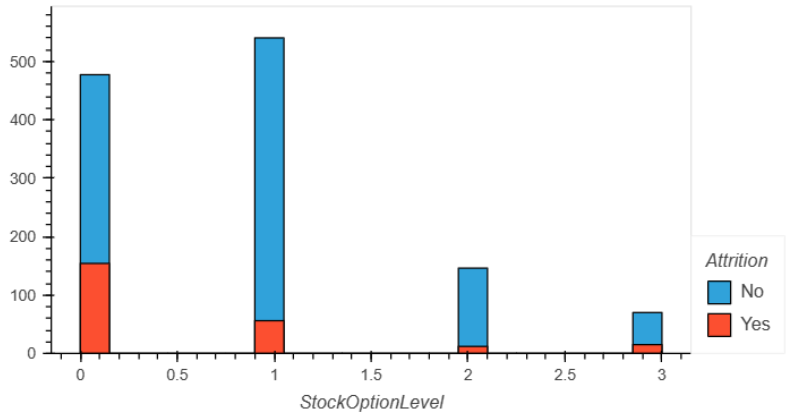
- This indicates that the data set is an imbalanced dataset where the number of observations belonging to class 1 (No) is significantly higher than those belonging to class 0 (True) .
- The conventional accuracy of the predictive models is not a relevant measure of model performance because machine learning algorithms are usually designed to improve accuracy by reducing the error. Thus, they do not take into account the class distribution / proportion or balance of classes.
- Hence, we will consider other model performance measures to evaluate a model, keeping in mind the class imbalance problem.

2. Employee Attrition based on Job Roles

- Majority of the employees are from the Research scientist department and sales.
- Research Scientist has the highest attrition rate- Could be a consequence of the higher number of people from Research.
- Overall Attrition is more in sales as it has a smaller number of employees and more attrition as compared with Research.
- Job roles of employees having maximum attrition are Sales Executive, Laboratory Technician and Research Scientist.



3.Employee Attrition based on Stock Option level.

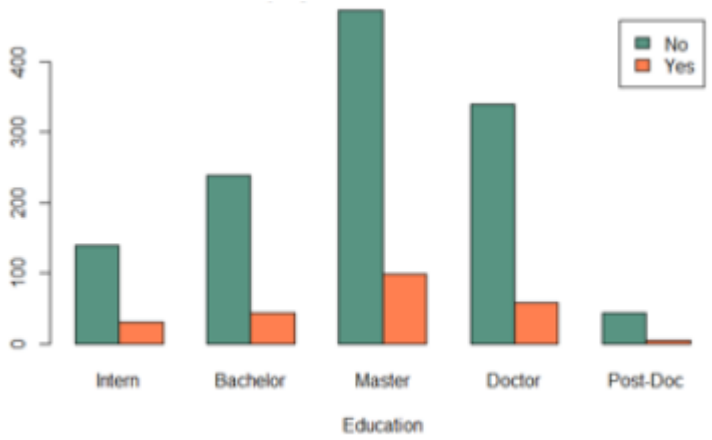


Higher Stock option level is inversely proportional to Attrition rate.employees with less stock option level leave the company more often.so providing more stock options could reduce the attrition rate.

4.Employee Attrition based on Business Travel

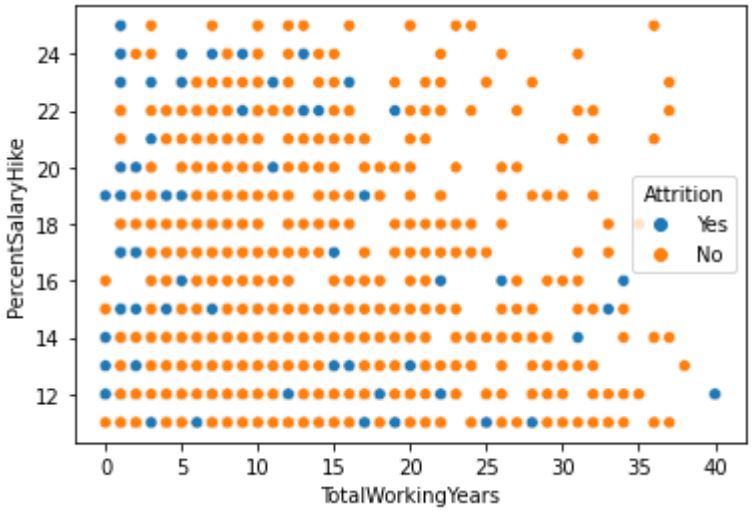
There are more people who travel rarely compared to people who travel frequently. In case of people who travel rarely, people have left the company and in other cases attrition rate doesn't vary significantly on travel. Since it has shown it has played an important role in a few cases, it is red an important variable.considered

5.Employee Attrition based on Education



When compared with Education level, we have observed that employees in the highest level of education in their field of study have left the company. We can conclude that the Education Field is a strong indicator of attrition.

6.Employee Percentage Salary Hike by Total Working Hours:

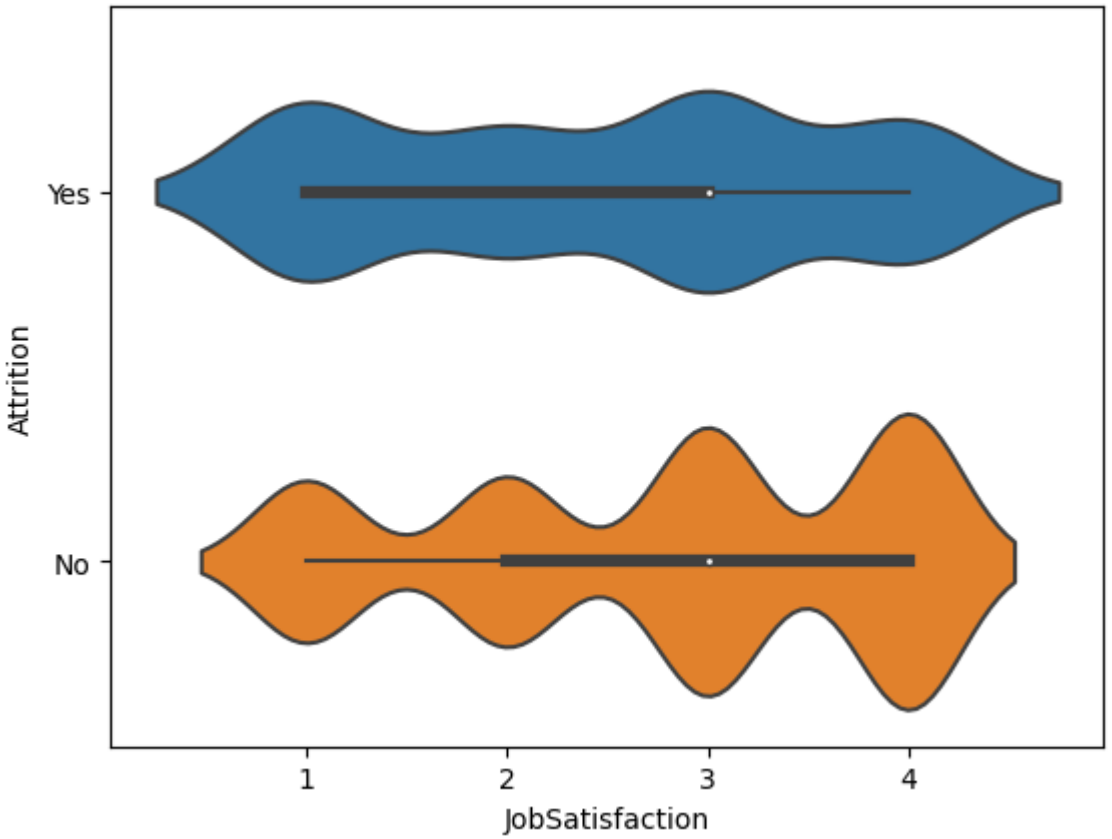


As we can see from the graph, there is no linear relationship between total working years and percentage of salary increases. People with fewer years of work experience have received a 25% salary increase when switching jobs,but this pattern does not exist for those with more years of experience. After 20 years, there is no attrition among those who received a 25% raise.

As the number of years of experience increases, those who switch over receive less than a 20% raise, as the trend indicates.

7.Employee Attrition by Job satisfaction:

Violin Plot is a technique for visualising the distribution of numerical data from various variables. It's similar to a Box Plot, but it has a rotated plot on each side that provides more information about the density estimate on the x-axis or y-axis.



.

The density of yes attrition lies between 1 to 3 ,and no attribute represents between 2 and 4.

There is a visible trend in the category of people who do not leave, but not in the category of people who do leave. People with high and very high job satisfaction are strongly represented in the attrition group.

STATISTICAL APPROACH:

In this project we will explore the basics of statistical analysis and why it is important to use it in subjects such as data science.

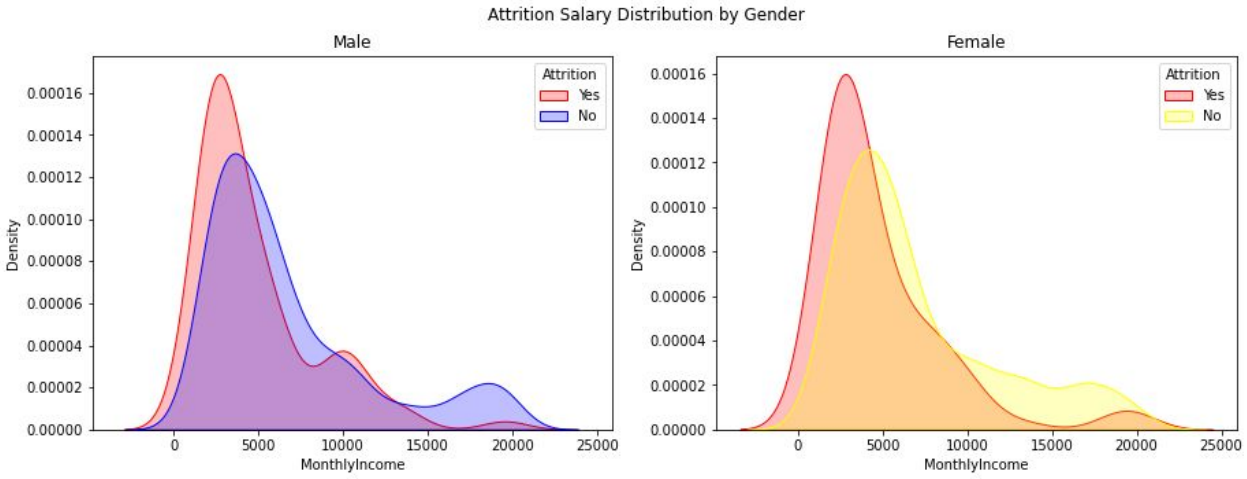
Normal Distribution:

Normal Distribution: Also known as bell curve, is a distribution in which half of the data lies on the left side and the other half lies on the right side of the distribution. In this distribution the curve is symmetric and the mean, mode and median are all equal.

Right Skewed Distribution:has a long tail pointing to the right. This means that in our sample or population most of the data is concentrated to the left side of the distribution. Left Skewed Distribution: has a long tail pointing to the left. This means that in our sample or population most of the data is concentrated to the right side of the distribution.

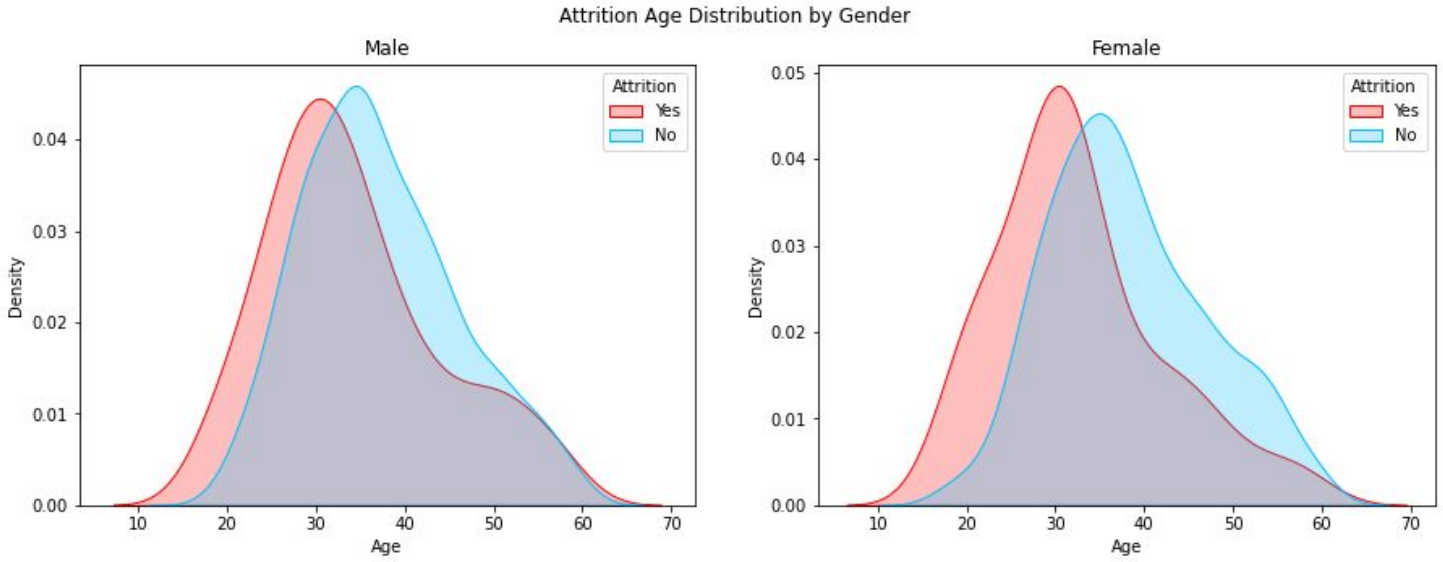
1.Attrition Salary Distribution By Gender

The below graph represents salary distribution by gender , the male and female employees have the same salary distribution.



The Salary distribution in the male and female Employees is similar

2.Attrition Age Distribution by Gender:



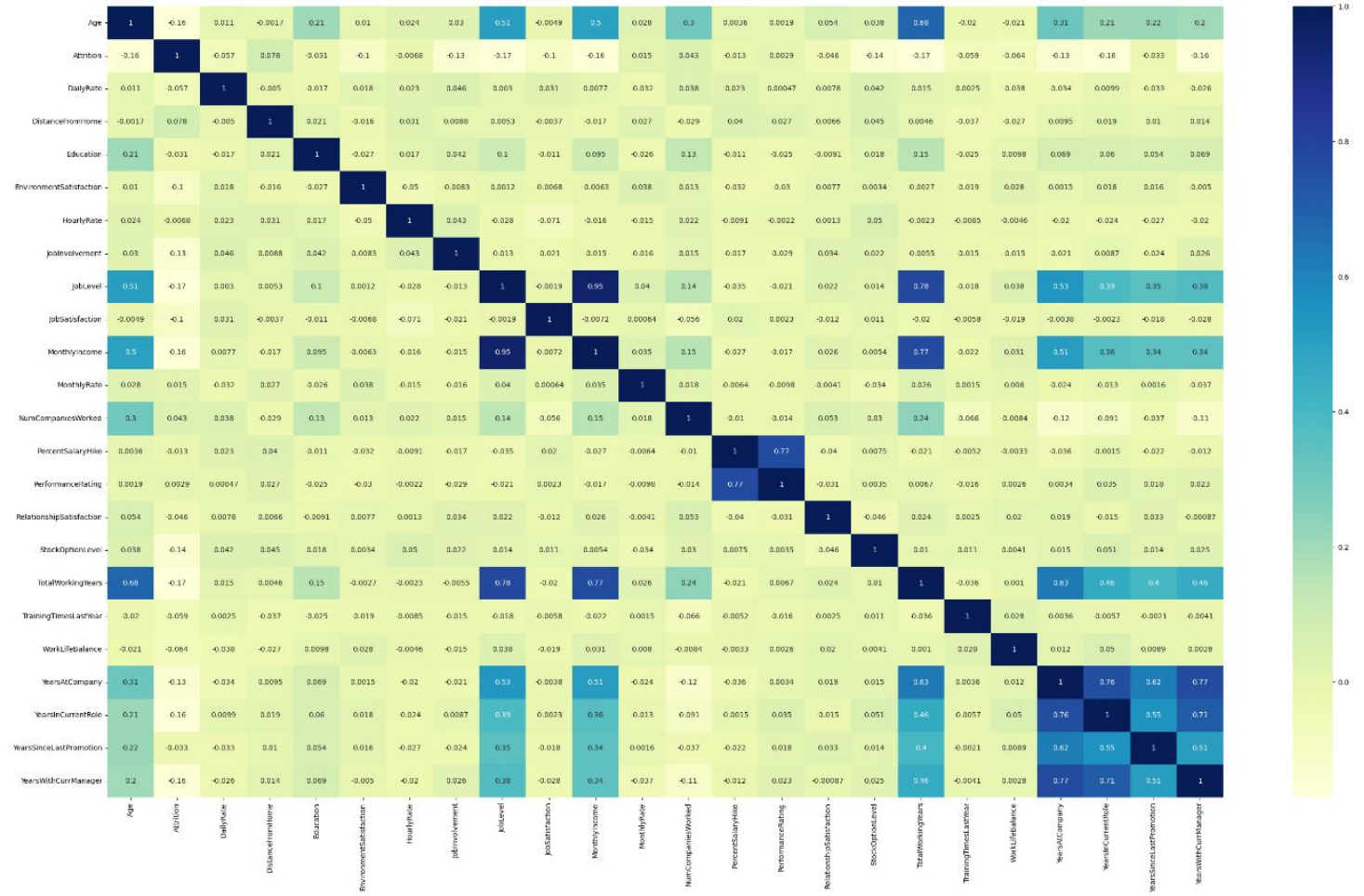
From the above distribution graph, Male Employees have maximum Attrition compared to Female Employees.

3.Correlation:

We can see from the correlation plots that many of our columns appear to be poorly correlated with one another. In general, when developing a predictive model, it is preferable to train a model with features that are not overly correlated with one another in order to avoid dealing with redundant features. If we have a large number of correlated features, we could use a technique like Principal Component Analysis (PCA) to reduce the feature space.

As we can see, the target column does not have a strong correlation with any of the numerical columns. However, other correlations can be found, such as;

- More senior employees have more total working years (very obvious)
- Higher performance ratings result in a greater percentage of salary increases.
- The more years an employee works, the more their monthly income grows.
- A lot of employees remain in their current role and also under the same manager as years pass by meaning they don't get promotion and this could be a major factor contributing to attrition.
- From here, we can deduce that the lack of promotions may be a Significant factor to attrition.


Age
Attrition
DailyRate
DistanceFromHome
Education
EnvironmentSatisfaction
HourlyRate
JobInvolvement
JobLevel
JobSatisfaction
MonthlyIncome
MonthlyRate
NumCompaniesWorked
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager
1.0
0.8
0.6
0.4
0.2
0.0

MODEL EVALUATION

After pre-processing, we split our data into training, and test dataset. From a total of 1470 observations,

we choose:

1. 70% observation for *Training Dataset.*
2. 30% observation for Test *Dataset.*

Evaluation with Logistic Regression

Logistic Regression statistical model (also known as *logit model*) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

Accuracy Score:0.8794104308390023

```
TESTING RESULTS:
================================
CONFUSION MATRIX:
[[358  13]
 [ 44  26]]
ACCURACY SCORE:
0.8707
CLASSIFICATION REPORT:
                   0.0         1.0  accuracy   macro avg  weighted avg
precision     0.890547    0.666667  0.870748    0.778607      0.855011
recall        0.964960    0.371429  0.870748    0.668194      0.870748
f1-score      0.926261    0.477064  0.870748    0.701663      0.854960
support     371.000000   70.000000  0.870748  441.000000    441.000000
```

Evaluation with Random Forest Classifier

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilises ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. It generates predictions without requiring many configurations in packages.

Accuracy Score:0.8258049886621315

```
TESTING RESULTS:
================================
CONFUSION MATRIX:
[[483   0]
 [105   0]]
ACCURACY SCORE:
0.8214
CLASSIFICATION REPORT:
                    0      1  accuracy   macro avg  weighted avg
precision    0.821429    0.0  0.821429    0.410714      0.674745
recall       1.000000    0.0  0.821429    0.500000      0.821429
f1-score     0.901961    0.0  0.821429    0.450980      0.740896
support    483.000000  105.0  0.821429  588.000000    588.000000
```

Model	Accuracy	Precision
Logistic Regression	85%	89%
Random Forest	82%	82%

Results

- ★ The data analysis and prediction from the logistic regression machine learning model gives an accuracy of 87%, and precision of 89% which shows that model is working efficiently.
- ★ The model gave an accuracy score of 82%, not too bad. The random forest works quite well even with the default parameters. That's one of the reasons we used RF for this problem.Though this can be improved by tuning hyper parameters of Random Forest classifier. Random forest also doesn't over fit easily because of its randomness feature.
- ★ Most of the employee job roles are Sales Executive, Research Scientist and Laboratory Technician
- ★ Job roles of employee having maximum attrition is Sales Executive, Sales Representative, Laboratory Technician and Research Scientist
- ★ Job Roles having least employee attrition are Research Director, Manager and Healthcare representative
- ★ The workers with low JobLevel, MonthlyIncome, YearAtCompany, and TotalWorkingYears are more likely to quit their jobs. BusinessTravel : The workers who travel a lot are more likely to quit than other employees.
- ★ Department : The workers in Research & Development are more likely to stay than the workers in other departments.
- ★ EducationField : The workers with Human Resources and Technical Degree are more likely to quit than employees from other fields of education.
- ★ Gender : The Male are more likely to quit.
- ★ JobRole : The workers in Laboratory Technician, Sales Representative, and Human Resources are more likely to quit the workers in other positions.
- ★ MaritalStatus : The workers who have Single marital status are more likely to quit than the Married, and Divorced.
- ★ OverTime : The workers who work more hours are likely to quit then others

Conclusion

We only have 1,470 observations from the company, which is insufficient to create a robust machine learning algorithm to predict wear in advance and act on this prediction. For that reason, we need to have more data from the company to improve the prediction accuracy of our model.

The improvement of the modelling depends more on the increase in the amount of observation. In other words, the company should focus more on collecting reasonable data from its employee.

Based on the above variables, one can clearly notice a pattern. The employees are more concerned with the materialistic objects that they get directly in hand. Then comes the psychological variables that determine if an employee might leave the organisation.

- As is well reflected in the model, employees who work overtime are much more likely to quit. Therefore, the company must understand the reason they are working overtime. Is it because of a workload too high or the qualifications of the employees are not sufficient to complete scheduled tasks on time? Maybe there could be some other reasons behind it. Our recommendation will be to understand the reason for the overtime with a detailed investigation and take appropriate action to reduce the factors behind this dropout factor.
- 5% of the employees work as a Sales Representative and 17% of the employees work as a Laboratory Technician. They have a percentage decrease of 39% and 17% respectively. These two-job roles must be questioned, and the company must find the reasons why these roles face a higher attrition rate than all others and take the necessary measures.
- 18% of employees travel frequently and have the highest percentage of attrition (25%). The company must ask itself what makes travelling a burden for your employees. The company must balance the status of the trips and, if it is necessary, there may be some adjustments to the job description in terms of travel. The company can use some additional incentives to motivate your employees who are supposed to travel.

The company should primarily seek to increase the effectiveness of these factors. As a result, it will yield to the decrease in the dropout rate.

Suggested Actions:

It's not sensible to focus on every employee who wants to leave because it costs time and energy for the human resources management department. HR department need to focus on:

- ➔ Improving the work conditions: Provide an option for the employee's to work from home, on a flexible schedule, or in an office with an ergonomic workspace, they will be more satisfied with their work and more likely to achieve a healthy work-life balance.
- ➔ Offer modest salaries and perks To maintain the critical employee's company need's to offer equitable and modest salaries. You can also give added perks like flexible schedules, travel discounts etc.

- ➔ Employee Engagement: When you have talented employees we need to find ways that you can help expand the employee's skill set, so that their involvement in the job increases. If their involvement is low, they will get bored and think that they are not growing within the organisation.

References

1.“What Is Employee Attrition? Definition, Attrition Rate, Factors, and Reduction Best Practices |.” *What Is Employee Attrition? Definition, Attrition Rate, Factors, and Reduction Best Practices | Spiceworks*, 27 May 2020, www.spiceworks.com/hr/engagement-retention/articles/what-is-attrition-complete-guide.

2.“IBM Employee Attrition Analysis.” *IBM Employee Attrition Analysis*, arxiv.org/pdf/2012.01286.pdf.

3.Et.al, Sadineni Sanjeetha. “Analysis of Employee Attrition Using Machine Learning Techniques.” *Turkish Journal of Computer and Mathematics Education (TURNCOAT)*, vol. 12, no. 6, Auricle Technologies, Pvt., Ltd., Apr. 2021, pp. 28–31. *Crossref*, https://doi.org/10.17762/turcomat.v12i6.1253.

4.Dallal, G. E. (1988, November). LOGISTIC: A Logistic Regression Program for the IBM PC. *The American Statistician*, *42*(4), 272. https://doi.org/10.2307/2685137

5.*sklearn.ensemble.RandomForestClassifier*. (n.d.). Scikit-learn. Retrieved January 26, 2023, from https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html