HR Analytics For Predicting Employee Attrition with Logistic Regression

Dr.Sailaja Nimmagadda[1]
Assistant Professor
Department of Business Management
Velagapudi Ramakrishna Siddhartha Engineering College,
Vijayawada, Andhra Pradesh
India-520007
Email: nsailaja@vrsiddhartha.ac.in
Dr.R.Jeya Lakshmi[2]
Assistant Professor
Master of Business Administration
Laki Reddy Bali Reddy College of Engineering,
Affiliated to JNTU Kakinada,
Mylavaram
India-521230
Mr.Surapaneni Ravi Kishan[3]
Associate Professor
Department of Computer Science and Engineering
Velagapudi Ramakrishna Siddhartha Engineering College,
Vijayawada, Andhra Pradesh
India-520007
Mrs. Naga Lakshmi veeram[4]
Assistant Professor
Andhra Loyola Institute of Engineering and Technology
Vijayawada, Andhra Pradesh
India

Abstract:
Today most of the organisations are failing to retain the valuable employees. As high Employee attrition rate can never be positive sign for an organization, it is necessary to identify, organize and manage the factors affecting Employee attrition. Present paper focused on developing an algorithm to predict employee attrition rate by considering the factors affecting employee attrition. Data considered for the study has been collected from secondary source Kaggle Data sets. Simple random sampling Technique is used. Data containing 35 different attributes or factors of 1,500 employees belonging to IT Companies were selected. Python is used as a tool for conducting Data analytics. An exploratory data analysis constructed to see the effect of every factor against employee attrition. Feature Engineering was performed to select the most influencing factors affecting Employee attrition rate. Feature selection gave the five most influencing factors driving towards employee attrition rate. Further a machine learning algorithm Logistic regression was applied as its accuracy rate is high as compared to other. Based on the predicted Employee attrition rate, we have calculated Employee turnover rate. Major factors affecting the Employee attrition rate are identified as Environment satisfaction, Job involvement, Job satisfaction, Relationship satisfaction, Work life balance. Employee turnover rate is calculated by predicting Employee attrition rate. Total turnover rate is 16.122%. There could be lot of factors or the variants that could affect employee attrition. Present paper helps to study how various significant factors influence the attrition of employees and what kind of working environment drive towards attrition.

Keywords: HR analytics, Prediction, Logistic regression, HR metrics, Employee attrition

1. Introduction and Related work:

Profit generation and organisational success will be the primary motive of any organisation. This can be achieved with the efforts of efficient human resources of an organisation [1].To meet the competitive advantage organisations are hunting for capable human resources. With this advent, every successful organisation has to look forward to retaining efficient people as their attrition rate is increasing day by day and this loss of employees will have a severe impact on the ability of organisation to succeed [2].

In general, there might be many personal or organisational factors that decide employee attrition[3].Some researchers focused on machine learning techniques to study organisational factors which lead to employee attrition[4].Some studies analysed the reasons why employee quit the company and suggested improving incentive structure and work environment as measures to retain employees[5]. Data mining techniques were applied to study employee turnover [6].

People's analytics with a combination of data integration, Insight generation, and Organizational and workforce analysis help to track employee activities for assessing employee performance. However HR functions need to know the most influencing factors leading to employee attrition. HR analytics tools like Python can develop Logistic Regression machine learning algorithms to predict the accurate attrition rate of employees [7].

Many studies have concentrated on classification and data sampling methods but the application of Feature engineering [8, 9] for identifying the most influencing factors affecting employee attrition rate was rarely studied. Hence for the present study feature selection method was adopted and identified the five most influencing factors among thirty-five different attributes that lead to predicting employee attrition and turnover rate.

2. Methodology

Data considered for the study has been collected from secondary source Kaggle Data sets. A simple random sampling Technique is used. Data containing 35 different attributes or factors of 1,500 employees belonging to IT Company were selected. Python is used as a tool for conducting Data analytics. An exploratory data analysis was constructed to see the effect of every factor on employee attrition. Feature Engineering was performed to select the most influencing factors affecting Employee attrition rate. Feature selection gave the 5 most influencing factors driving towards employee attrition rate. Further, a machine learning algorithm Logistic regression was applied as its accuracy rate is high as compared to others. Based on the predicted Employee attrition rate, we have calculated the Employee turnover rate. The data set contains different features with descriptions related to employees as given in Table 1.

2.1 Exploratory Data Analysis:

To perform prediction, we need to understand the data and find insights based on analysis then perform an algorithm to predict employee attrition rate. To understand the data we need to perform exploratory data analysis and determine the hidden insights present in the data and how these are effecting employee attrition.

Table1: Variable Features and data types description

Variable	Datatype	Description
Age	Numeric	Age of Employee
Attrition	Text	employee attrition
BusinessTravel	Text	Journey specifically taken for work purposes
DailyRate	Numeric	Pay for an individual service for a single day
Department	Text	Department in which employee works
DistanceFromHome	Numeric	Distance from home to work place
Education	Numeric	Education ualification of employee
EducationField	Text	Background of education field
EmployeeCount	Numeric	count of employee in each dept
EmployeeNumber	Numeric	Unique number of every employee
EnvironmentSatisfaction	Numeric	Working environment satisfaction
Gender	Text	Gender of Employee
HourlyRate	Numeric	Pay for an individual service for one hour
JobInvolvement	Numeric	Level of Employee involvement in Job
JobLevel	Numeric	Level of Employee in organization
JobRole	Text	Role of Employee in organization
JobSatisfaction	Numeric	Job satisfaction of Employees
MaritalStatus	Text	Marital Status of an Employee
MonthlyIncome	Numeric	Pay for an individual service for one month
NumCompaniesWorked	Numeric	Number of companies Employee worked earlier
Over18	Text	Age greater than 18
OverTime	Text	Employee working over time
PercentSalaryHike	Numeric	percentage of salary hike employee received
PerformanceRating	Numeric	Rating of employee performance
RelationshipSatisfaction	Numeric	Explains about work environment
StandardHours	Numeric	Number of standard hours
TotalWorkingYears	Numeric	Total number of working years of employee
TrainingTimesLastYear	Numeric	Number of training sessions attended in last year
WorkLifeBalance	Numeric	Wor life balance of Employee
YearsAtCompany	Numeric	Number of years worked with current company
YearsInCurrentRole	Numeric	Number of years worked in current role with current company
YearsSinceLastPromotion	Numeric	Number of years since last promotion of employee
YearsWithCurrManager	Numeric	Number of years with current manager

Portfolio Report: Number of observations, number of variables, total size in memory, total missing (%), Variable types.

Fig1: Portfolio Report

Portfolio report is one of the library in python which can be useful to analyze each feature and gives the report which contains Distinct count, missing(%),unique, infinite(%), missing(n), mean, infinite(n), minimum, maximum, zeros and other data based on feature.

2.2 Correlation Matrix:

The correlation matrix explains the dependency between the attributes. We can able to see the dependencies between the features by using a heat map. In our data the different features like age, daily rate, and distance from home, education etc. are correlated.
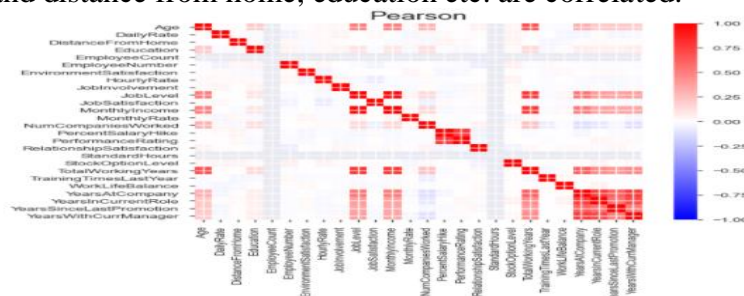

Fig2: correlation matrix

2.3 Visualizations of categorical features: Predicting employee attrition we need to

understand the data and find insights in the data so that Data contains different features these

are categorized as categorical and numerical. The features of categorical type are business travel, department, education, environmental satisfaction, Gender, and Job Role.


Fig 3: Relation between Employee Attrition and Business Travel, Education, Department, Environment satisfaction, Marital Status, Gender, Job role and Job involvement
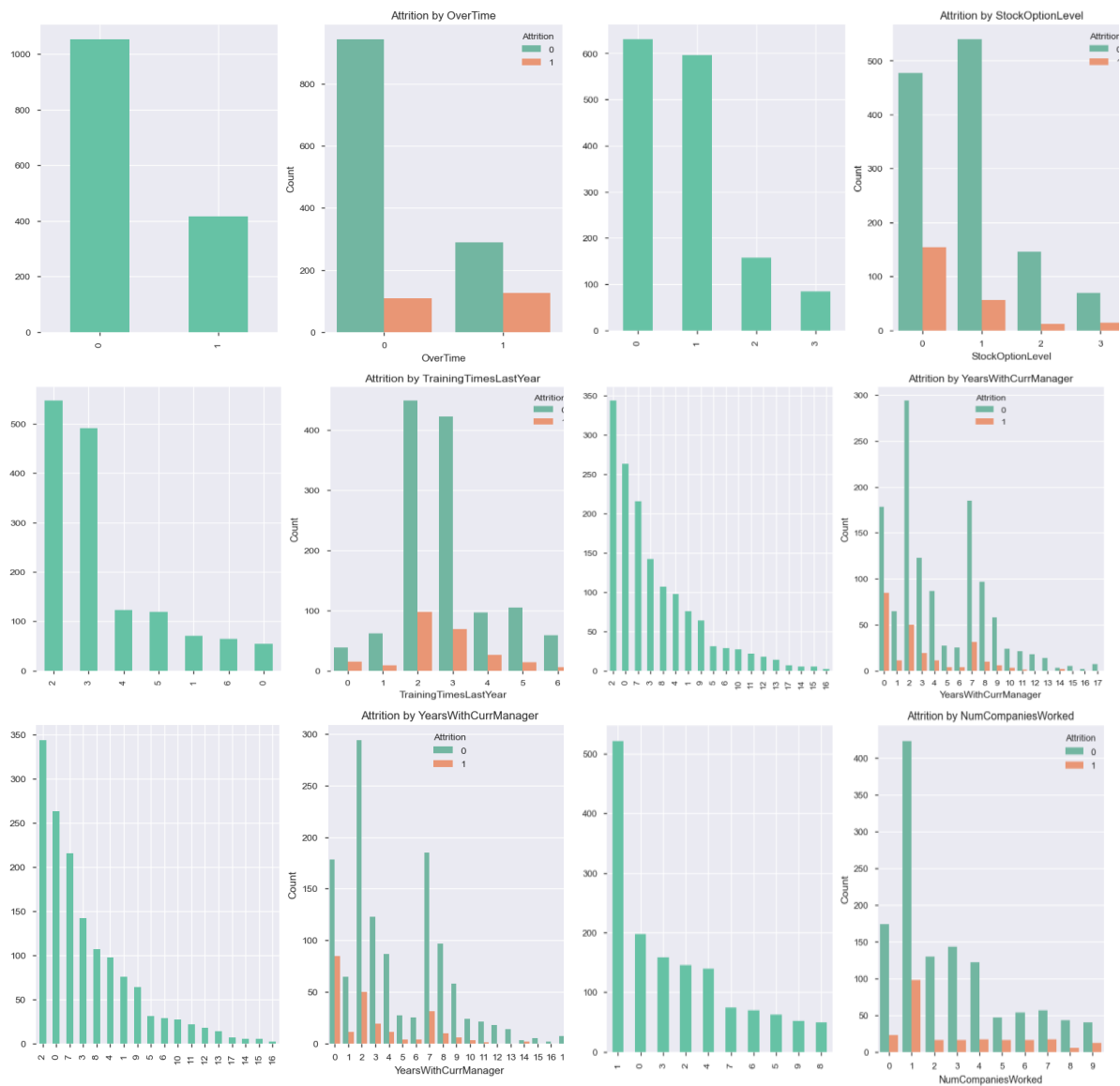
Fig 4: Relation between Employee Attrition and Over time, Stock level option, Training time, current Manager, Education field and No. of Companies worked.

2.4 Visualization of numerical features:

To apply the algorithm, we have analyzed numerical features data in our data set with employee attrition rate. We have analysed employee attrition with different factors like Age, Daily rate, Hourly rate, distance from home, Monthly income, job involvement, percentage of salary hike, total experience, years of service at the company, and service in a current role, since with last promotion, years with current Manager.
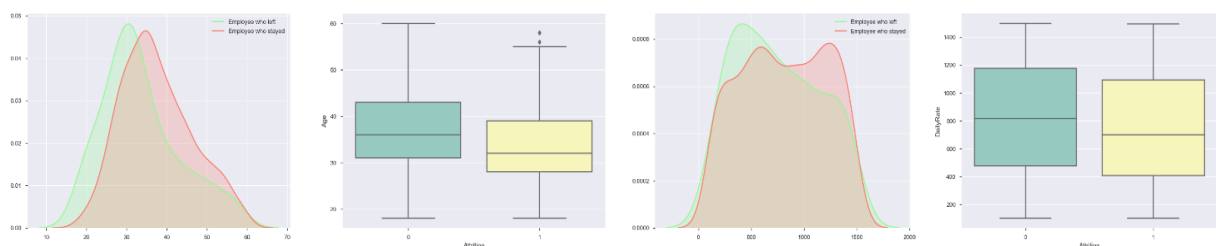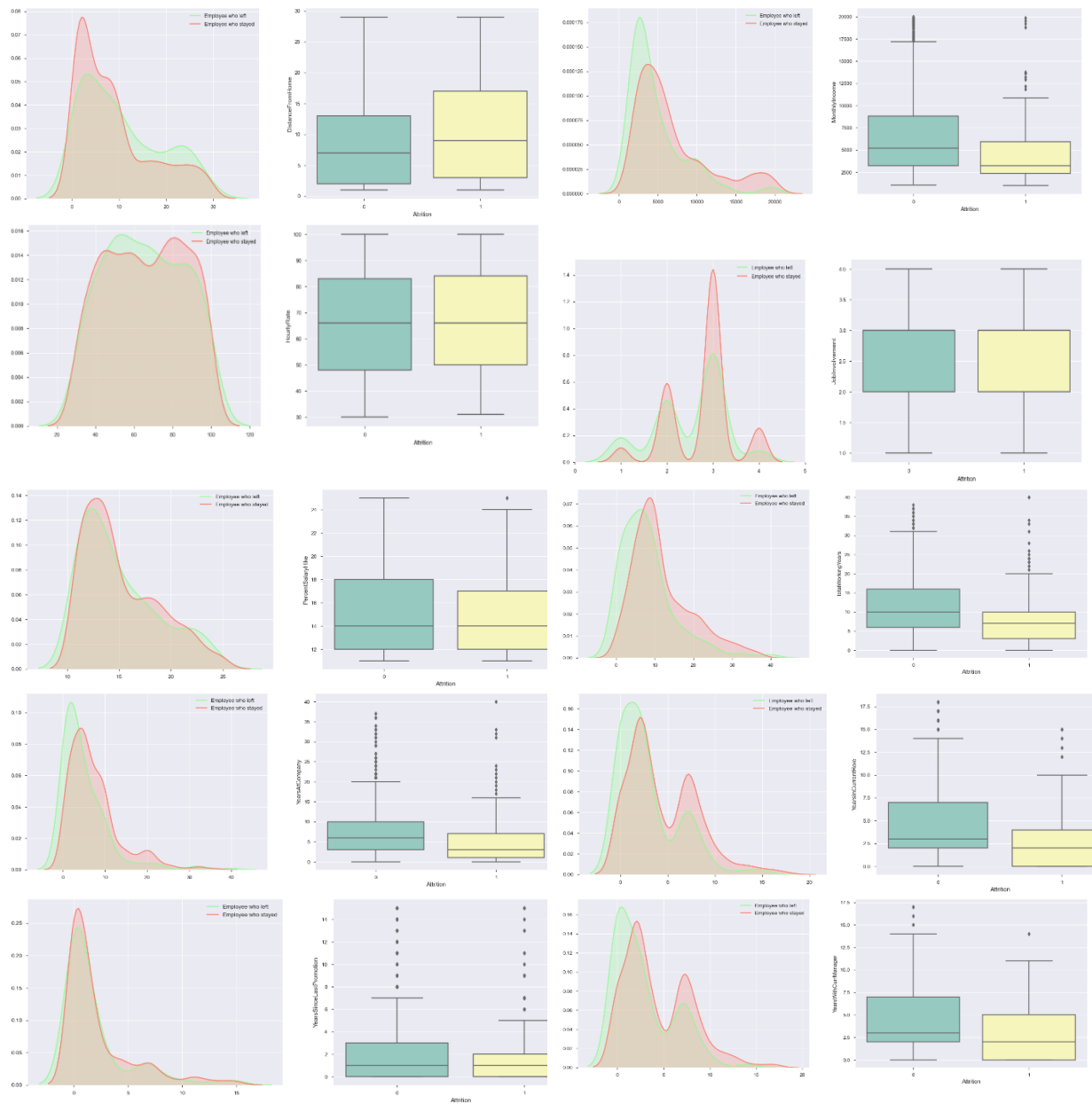
Fig 5: Employee attrition rate w.r.t Age, Daily rate, Hourly rate, distance from home, Monthly income, job involvement, percentage of salary hike, total experience, years of service at company, service in current role, since with last promotion, years with current Manager.

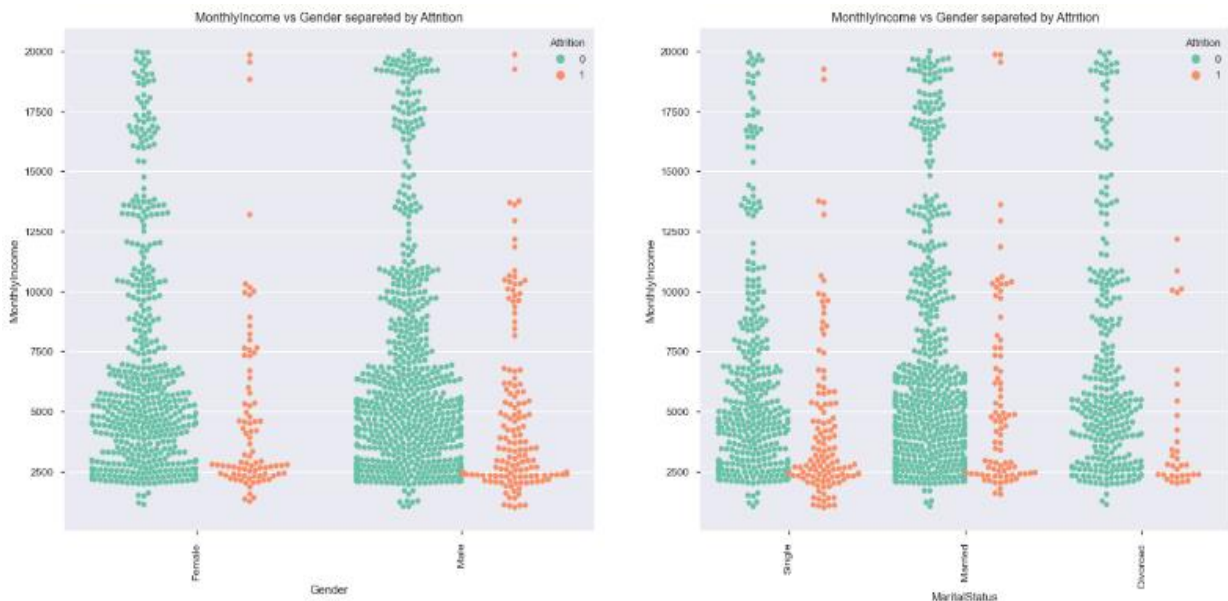## 2.5 Visualization of Categorical vs Numerical Features

In next step we are analyzing both categorical and numerical features
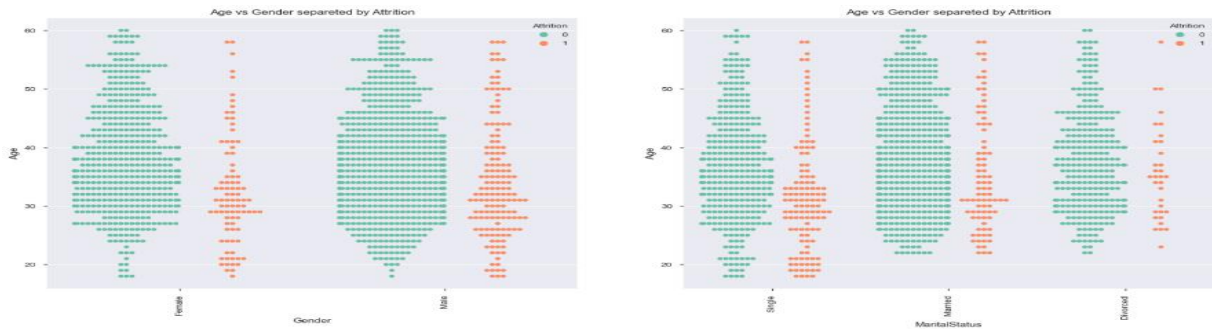
Fig 6: Age vs Gender and Marital Status w.r.t Attrition Rate
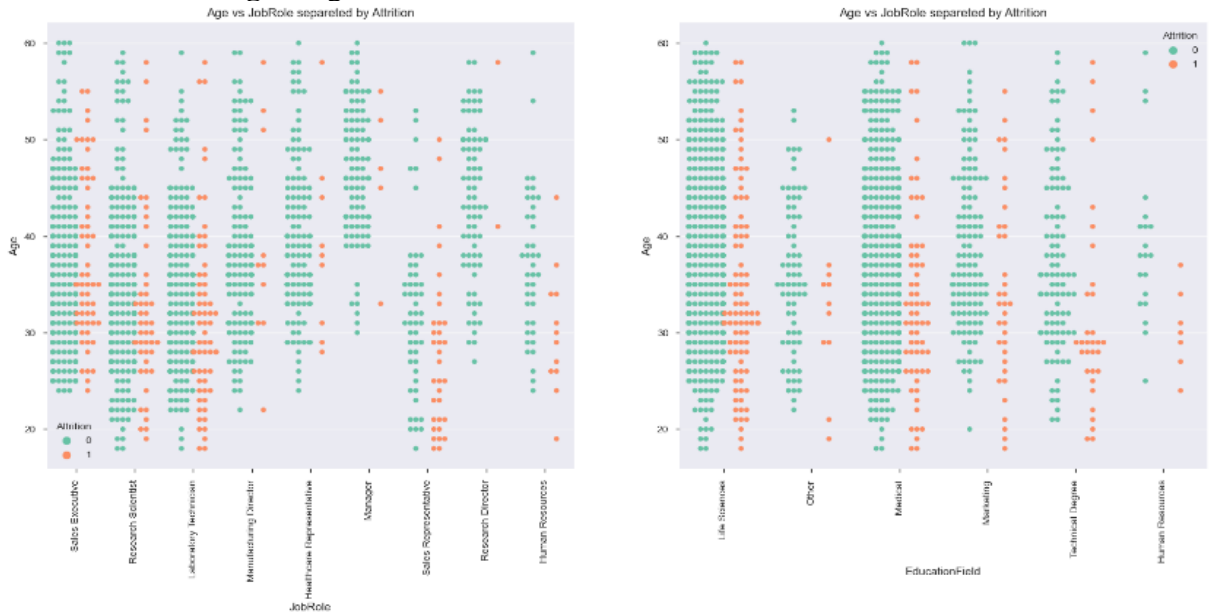
Fig 7: Age vs Job Role and Education Field w.r.t Attrition Rate

Fig 8: Monthly Income Vs gender and marital status w.r.t Attrition Rate

3. Results and Discussion

Feature Engineering

After analyzing all the selected features that affect the employee attrition rate i.e. total satisfaction, five features namely Environment Satisfaction, Job Involvement, Job Satisfaction, Relationship Satisfaction, and Work-Life Balance are selected, as these are identified as the most influencing factors that affect the employee attrition rate.

df['Total Satisfaction'] = (df['Environment Satisfaction'] + df['Job Involvement'] + df['Job Satisfaction'] + df['Relationship Satisfaction'] df['Work Life Balance']) /5
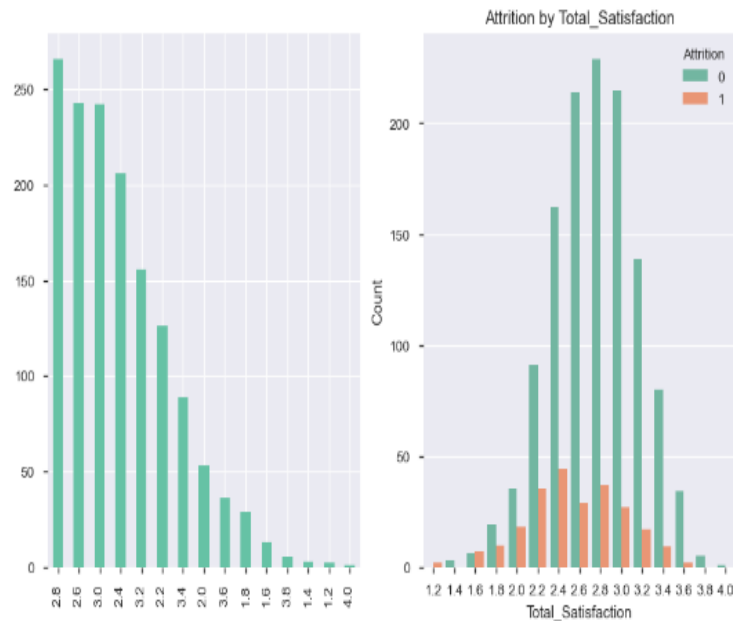

Fig 9: Total Satisfaction

Machine Learning Algorithms:

To predict employee attrition, we have to apply algorithms based on our data set for better results. In present context, researcher has selected Logistic Regression.

Logistic Regression:

Logistic Regression is used in statistics as a model of mathematics to approximate the possibility of occurrence of events based on past statistics. Binary data is the mechanism used in it. If an occurrence takes place it is considered as "1" and the event that does not happen is considered as "0".

Output:

HR Metrics Calculation:

In present data set employee attrition rate is calculated for target column.

Number of employees left the Organization

Turnover rate = _____

Total number of Employees during the year

Turnover Rate= [237/1470]*100 = 16.122%

Therefore, Total turnover rate in a given year is 16.122%

4. Conclusion and Future scope:

Employee and organization's investment in employees is very valuable; therefore it is advisable to reduce the Employee attrition rate as much as possible. Employee attrition could be due to many factors as explained in the present paper. The present study aims to predict employee attrition and develops an algorithm to predict employee attrition rate using the logistic regression method with feature selection. Prediction is made by considering the factors that affect employee attrition. In the data set the interrelation between every attribute to attrition rate is analysed and the most impacting attributes were identified for overall satisfaction as work environment, job involvement, job satisfaction, relationship satisfaction and work-life

balance. These were selected and considered for the Feature engineering prediction process. Various algorithms like decision tree, logistic regression, and linear regression were performed and among all algorithms, logistic regression suits as the best algorithm as it gives high accuracy in both true and false scenarios. The paper also highlights the KPI- Key performance indicator - Turnover rate which is the most used HR metric. From the data set Total turnover Rate in a given year can be declared as 16.122%.Attrition cannot be removed but can be reduced hence companies need to concentrate on measures to improve the work environment, job involvement, job satisfaction, relationship satisfaction and work-life balance.

Nevertheless, there is a scope for improvement of the study by trying several sampling methods to evaluate the results and to find an optimum sampling technique for the study. Several retention strategies for employees can be developed based on employee turnover factors identified by the study. Further neural networks can be used to observe datasets and predictions. Furthermore, a cloud platform through a web interface can be developed which helps organisations in predicting employee attrition.

5. References:

[1] Chanodkar A, Changle R, Mamtani D. Prediction of employee turnover in organizations using machine learning algorithms. Prestige International Journal of Management and Research. 2019;12(1/2):222-6.

[2] Pratibha G, Hegde NP. HR Analytics: Early Prediction of Employee Attrition using KPCA and Adaptive K-means based Logistic Regression. In2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS) 2022 May 9 (pp. 11-16). IEEE.https://doi.org/10.1109/ICPS55917.2022.00010

[3] Harvida DA, Wijaya C. Faktor Yang Mempengaruhi Turnover KaryawandanStrategiRetensiSebagaiPencegahan Turnover Karyawan: SebuahTinjauanLiteratur. JIANA (JurnalIlmuAdministrasi Negara). 2020;18(2):13-23.

[4] Raza A, Munir K, Almutairi M, Younas F, Fareed MM. Predicting employee attrition using machine learning approaches. Applied Sciences. 2022 Jun 24;12(13):6424.https://doi.org/10.3390/app12136424

[5] Fallucchi F, Coladangelo M, Giuliano R, William De Luca E. Predicting employee attrition using machine learning techniques. Computers. 2020 Nov 3;9(4):86.https://doi.org/10.3390/computers9040086

[6] Ghazi A, Ismail Elsayed S, ElsayedKhedr A. A proposed model for predicting employee turnover of information technology specialists using data mining techniques.International journal of electrical and computer engineering systems. 2021 Jun 21;12(2):113-21.. https://doi.org/10.32985/IJECES.12.2.6

[7] Setiawan IA, Suprihanto S, Nugraha AC, Hutahaean J. HR analytics: Employee attrition analysis using logistic regression. InIOP Conference Series: Materials Science and Engineering 2020 Apr 1 (Vol. 830, No. 3, p. 032001). IOP Publishing.https://doi.org/10.1088/1757-899X/830/3/032001

[8] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018 Jul 26;300:70-9.9. https://doi.org/10.1016/j.neucom.2017.11.077

[9] Wardhani FH, Lhaksmana KM. Predicting Employee Attrition Using Logistic Regression with Feature Selection. Sinkron: jurnaldanpenelitianteknikinformatika. 2022 Oct 3;7(4):2214-22.https://doi.org/10.33395/sinkron.v7i4.11783

[10]Pavansubhash. IBM HR Analytics Employee Attrition & Performance.Kaggle. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attritiondataset? Select=WA_Fn-UseC_- HR-Employee-Attrition.csv, 2017.