Technical Coding Research Innovation, Navi Mumbai,
Maharashtra, India-410206

(HR Employee Attrition)

A Case-Study Submitted for the requirement of
Technical Coding Research Innovation
For
the Internship Project work done during

DATA SCIENCE WITH MACHINE LEARNING AND PYTHON

by
Krishna Desai
Kiarah Patel
Surabhi Nirgudwar
Akanksha jadhao

Date:
Grade:

Rutuja Doiphode
CO-FOUNDER &CEO
TCR innovation.

Name – Surabhi Nirgudwar
Internship Program – Data Science with Machine Learning and Python
Batch – January 2022
Certificate Code – TCRIB1R94
Date of submission – 11/05/2022

Abstract: Employees are an organization's most precious asset. They are the ones who provide value to the company in terms of quantity and quality. As a result, it is critical to keep a stable and promising workforce, which has become a difficult burden for employers over time, resulting in rising attrition in firms. This study article aims to investigate the causes of attrition from several perspectives

Keywords: Attrition, employee, employer, leadership, management, productivity, retention, organization.

I. INTRODUCTION

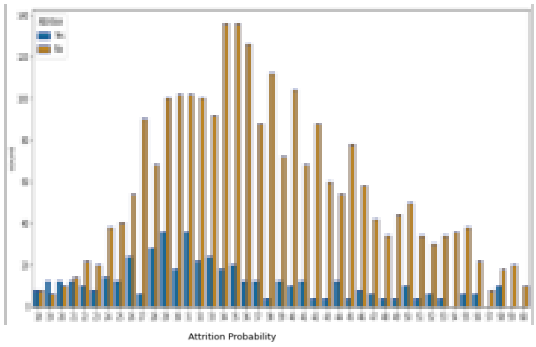
Attrition refers to the loss of a workforce for whatever reason. Attrition is a prevalent problem in any organization, regardless of the type of industry or the structure of the organization. It not only impedes output but also results in high long-run expenses and a loss of goodwill. As a result, there is a pressing need to look into this multi-faceted problem and find workable answers.

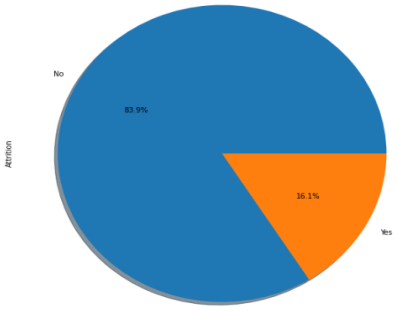
II. Exploratory data analysis

Exploratory Data Analysis (EDA) is a method of analyzing data through the use of visual techniques. With the use of statistical summaries and graphical representations, it is used to find trends, patterns, also examine assumptions. The "HR EMPLOYEE ATTRITION DATASET" includes information about an employee's gender, age, business travel, department, education, relationship satisfaction, and other factors.

```
In [7]: df.columns
Out[7]: Index(['EmployeeNumber', 'Attrition', 'Age', 'BusinessTravel', 'DailyRate',
               'Department', 'DistanceFromHome', 'Education', 'EducationField',
               'EmployeeCount', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate',
               'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
               'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked',
               'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
               'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
               'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',
               'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
               'YearsWithCurrManager'],
              dtype='object')
```

The dataset consists of exactly 2940 employees' data, with each employee having 34 unique attributes.

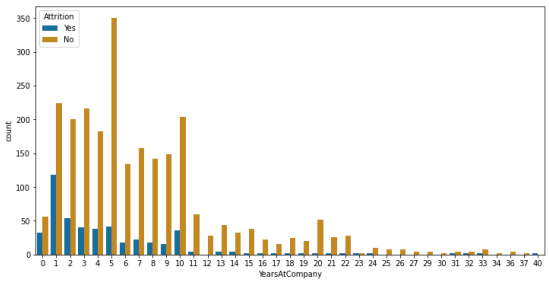




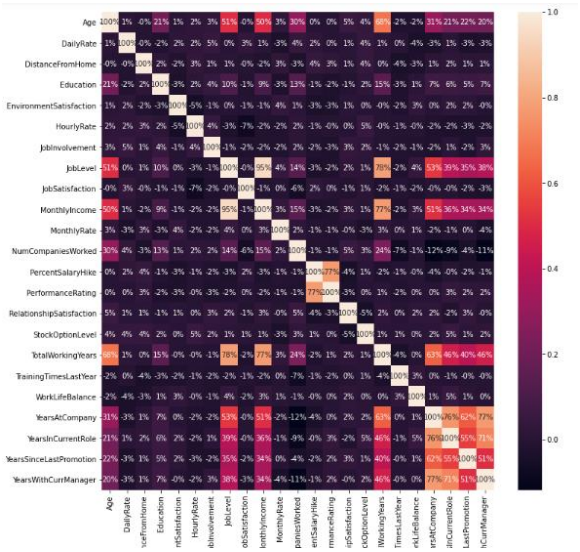
There are both numerical and category data in the dataset. In the given dataset there are neither missing values nor NAN values.

A thorough research was done to check if data there are extreme values and normalization was done accordingly

Name – Surabhi Nirgudwar
Internship Program – Data Science with Machine Learning and Python
Batch – January 2022
Certificate Code – TCRIB1R94
Date of submission – 11/05/2022



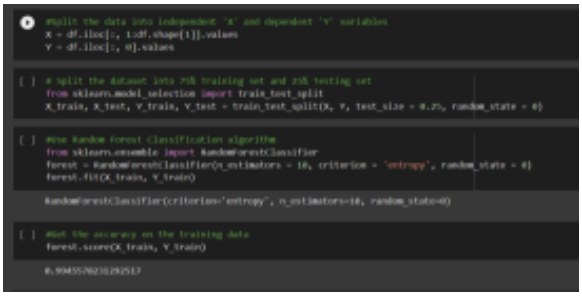
The above graph shows how a feature (years at company) is affecting the attrition rate. A correlation matrix is simply a table which displays the correlation. It is best used in variables that demonstrate a linear relationship between each other. The matrix depicts the correlation between all the possible pairs of values in a table.



III. Training and prediction of data

After pre-processing and EDA, Random Forest Classification approach to train the machine learning model because it can handle a high number of characteristics effectively. To begin, we divided the dataset into two parts: 80 percent for training and 20% for testing. which is loaded from the package of sklearn datasets.

After applying algorithm we using accuracy score to check the percentage of prediction.



```
In [39]: #Check model accuracy
         from sklearn.metrics import confusion_matrix
         cm=confusion_matrix(y_test,forest.predict(x_test))

         TN=cm[0][0]
         TP=cm[1][1]
         FN=cm[1][0]
         FP=cm[0][1]

         print(cm)
         print('Model Testing Accuracy is:',(TP+TN)/(TP+TN+FP+FN))

         [[488   0]
          [ 28  72]]
         Model Testing Accuracy is: 0.9523809523809523
```

IV. Conclusion

Attrition is unavoidable; it will happen; the only thing you can do is reduce it. When it comes to preventing attrition, intrinsic variables are just as significant as extrinsic factors, if not more so. Effective leadership can assist to control attrition to a large extent. Attrition does not always result in a negative impact on the business.

V. References

https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec

https://www.academia.edu/Documents/in/Employee_Attrition

Name – Surabhi Nirgudwar
Internship Program – Data Science with Machine Learning and Python
Batch – January 2022
Certificate Code – TCRIB1R94
Date of submission – 11/05/2022

https://www.researchgate.net/publication/307546344_EMPLOYEE_ATTRITION_AND_STRATEGIC_RETENTION_CHALLENGES_IN_INDIAN_MANUFACTURING_INDUSTRIES_A_CASE_STUDY