

HR Analytics Project- Machine Learning to Understand & Predict HR Attrition

**By – Shantanu Jagdish Jagtap (DS2402)**



**Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?**

*HR Analytics Project- Machine Learning to Understand & Predict HR*

*2*

*Attrition*



## Problem Definition:

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition 3*

Continuous technological innovation and evolving business models are key to the success of any enterprise in today's competitive business landscape. The ability to attract and retain top talent is the fundamental cornerstone of a successful business. Talented minds bring innovative ideas and solutions, transforming them into valuable products.

The key to success is effectively managing human resources.

The HR department is involved in the process. It invests in the continuous development of existing staff, as well as recruits and trains new employees. HR analytics involves gathering and analyzing data to enhance internal processes and inform critical decision-making. Employee attrition, the gradual loss of employees over time, is a significant challenge that HR analytics addresses. Retirement, voluntary resignation, or other reasons can lead to attrition.

High employee attrition is costly to an organization. There are expenses associated with job postings, hiring processes, paperwork, and training new hires. Losing talented employees can mean the loss of intellectual property.

Customer relationships may suffer in sales and marketing as

clients prefer to interact with familiar representatives.

Several factors contribute to employee attrition, including poor work-life balance, better job opportunities, salary discrepancies, lack of growth or recognition, and unhealthy relationships with managers. By utilizing machine learning classification techniques, organizations can examine these factors to forecast the likelihood of an employee leaving.

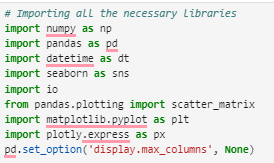
This project aims to investigate the main factors that contribute to employee attrition and develop predictive models using the IBM hr analytics employee attrition & performance dataset. This fictional dataset, developed by IBM and accessible on platforms like GitHub and Kaggle, comprises 1,470 records and 35 features that provide insights into employee characteristics. The main goals are to pinpoint the primary factors that contribute to attrition and create a predictive model using machine learning techniques.

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition 4*



Importing libraries for Data Analysis

To initiate the exploratory data analysis (EDA), we imported crucial libraries like pandas, numpy, seaborn, matplotlib, and others to facilitate data manipulation and visualization..



## Required Libraries



*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition 5*



Data Preparation: Load, Clean and Format

* We started by analyzing the dataset's structure and data types, finding that it consists of nine features, with eight of them being numeric and the remaining one being object-based. Among these numerical features, there are several ordinal variables, each with a distinct label corresponding to its numeric value. For example, the 'education' feature is encoded from 1 to 5, signifying levels ranging from 'below college' to 'doctorate'.





*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***6***

# Data Preparation: Load, Clean and Format



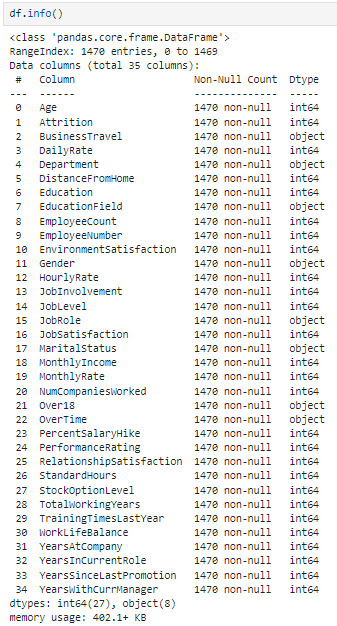
**Checking different datatypes in dataset:-**

**The ordinal features and their corresponding label encodings are as follows:**

* Education: 1 - 'Below College', 2 - 'College', 3 - 'Bachelor', 4 - 'Master', 5 - 'Doctor'
* Environment Satisfaction: 1 - 'Low', 2 - 'Medium', 3 - 'High', 4 - 'Very High'
* Job Involvement: 1 - 'Low', 2 - 'Medium', 3 - 'High', 4 - 'Very High'
* Job Satisfaction: 1 - 'Low', 2 - 'Medium', 3 - 'High', 4 - 'Very High'
* Performance Rating: 1 - 'Low', 2 - 'Average', 3 - 'Good', 4 - 'Excellent', 5 - 'Outstanding'
* Relationship Satisfaction: 1 - 'Low', 2 - 'Medium', 3 - 'High', 4 - 'Very High'
* Work-Life Balance: 1 - 'Bad', 2 - 'Good', 3 - 'Better', 4 - 'Best’

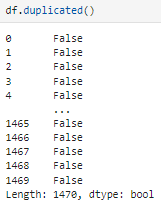
These labels are instrumental in interpreting the data, especially during EDA.

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***7***



# Data Preparation: Load, Clean and Format

**We also conducted a data integrity check to ensure the dataset's quality.**



**Fortunately, the dataset had no missing values, duplicate entries, or extraneous whitespace characters, which simplifies our analysis.**

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***8***

# Data Preparation: Load, Clean and Format



**Next, we examined the statistical summary of the dataset to gain insights. Some key observations include:**

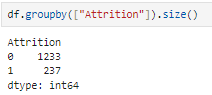
* The minimum employee age is 18, while the maximum is 60.
* The average distance from home to office is approximately 9.1 KM, indicating a typical daily round trip of at least 18 KM.
* The average performance rating is 3.163, with a minimum of 3.0, suggesting that most employees are rated as 'Good'. Notably, attrition among employees with an 'Outstanding' rating of 5 warrants further investigation.
* Half of the employees have worked for at least two companies previously.
* The presence of outliers in features such as Monthly Income and Monthly Rate is evident from their statistical distribution.
* Some features, as indicated by discrepancies between their mean and median values, exhibit skewness.
* For ordinal features, metrics like mean, median, and standard deviation are not as informative due to their categorical natur e.
* The features 'Standard Hours' and 'Employee Count' contain a single unique value across all records, making them non-informative.
* These preliminary analyses set the foundation for more detailed exploration and modeling**.**

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***9***



# Exploratory Data Analysis(EDA)

**Exploratory data analysis is the process of making initial inquiries about data to determine patterns, find irregularities, test hypotheses and assess assumptions through using summary statistics and graphical methods.**

**Data exploration of Target variable. ** **83.88% (1237 employees) Employees that left the organization while**

**16.12%(2379 employees )did leave made this dataset is considered to be**

**imbalanced because a lot of people stay in the company than they actually leave it.**

**Let say in this dataset we have education, department,education field, job role and also within these feature there are interrelated features such as Job satisfaction which is a among the value extract from response of all above answer. Mismatch between job role & job position can cause attrition. Let explore this feature by drawing all these features one after the other for more information.**

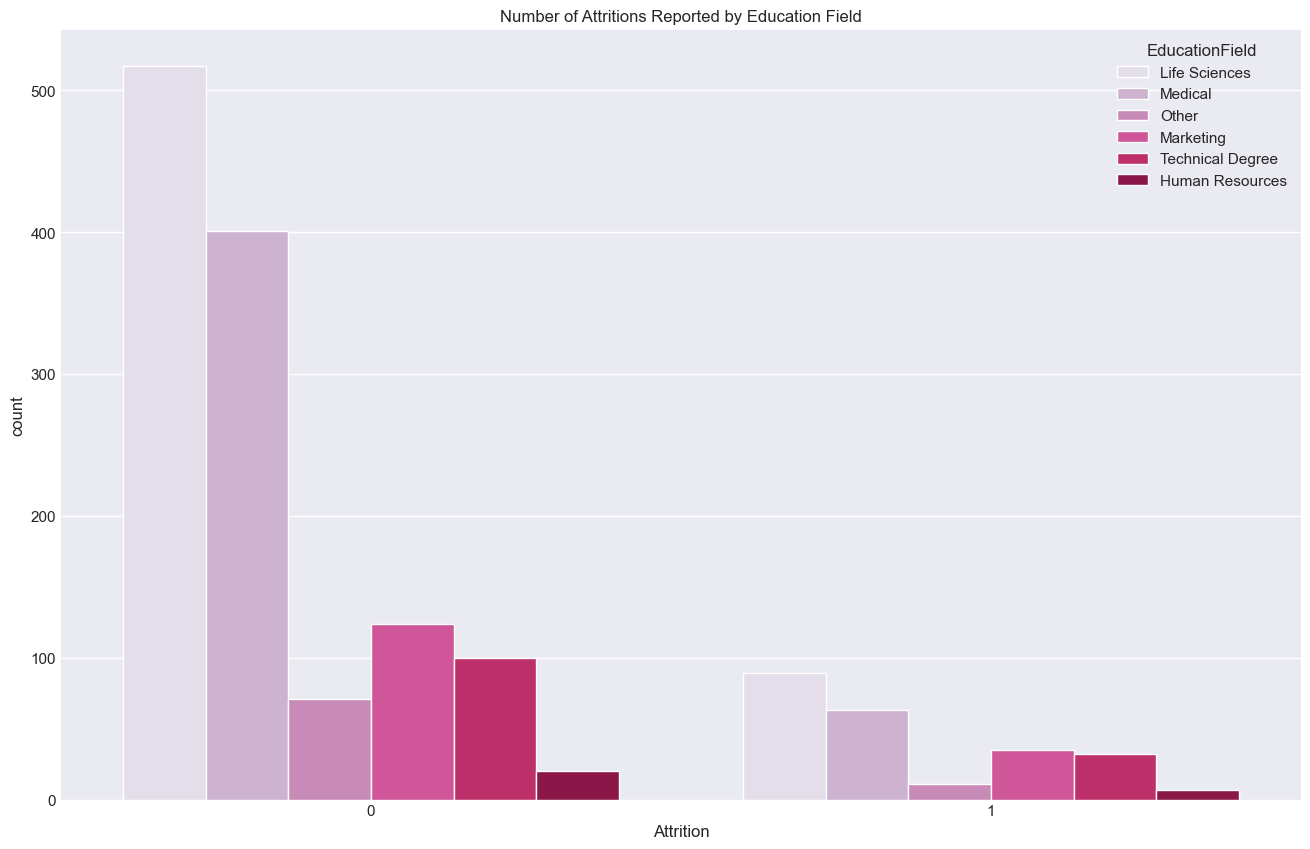
*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***10***

# Exploratory Data Analysis(EDA)



**Education level -**

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***11***



# Exploratory Data Analysis(EDA)

**Main Lessons from the Bar Plot:**

**High Retention in Science and Medical Jobs:- Most of the workers who did not leave (attrition = 0) are working in science and medical areas which means that these fields have higher than average retention rates.**

**High Workforce Turnover across Different Professions:- This shows the highest attrition rates of about 30% among those who left (attrition = 1), Medical, Life Sciences and Technical Degree fields even though they have a high retention rate.**

**Reasonable Levels of Attrition for Technical Degrees and Marketing:- These data also show moderate levels of attrition in technical degrees and marketing careers, making up nearly 19% of individuals who didn’t leave their jobs.**

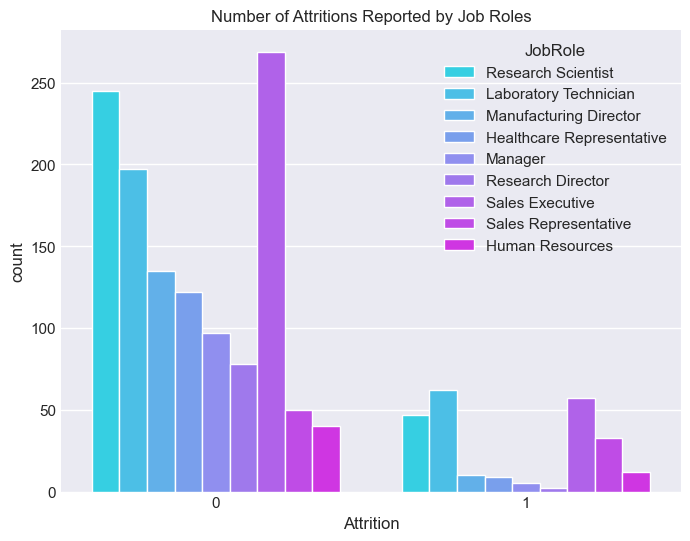
**Scarce Staffing of Human Resources Department:- Human resources accounts for just over 12 percent of employees, with low counts in both attrition and no attrition cases, implying a smaller workforce concerning this aspect.**

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***12***

# Exploratory Data Analysis(EDA)



**Number of attritions reported by different job roles-**



*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***13***



# Exploratory Data Analysis(EDA)

**Bar Plot Key Insights:**

**High retention rate in sales executives and research scientists:- More than 35% of attrition=0 employees are Sales Executives and Research Scientists, indicating that these job roles have a higher retention.**

**Moderate attrition across research scientist and sales representative roles:- Close to 30% of attrition=1 employees belong to the group Research Scientists, Sales Representatives, which shows that for these positions there are moderate attrition rates.**

**Large number of healthcare representatives and laboratory technicians:- Almost 19% of employees who did not leave are Healthcare Representatives and Laboratory Technicians, showing a substantial proportion in this position with reduced staff turnover.**

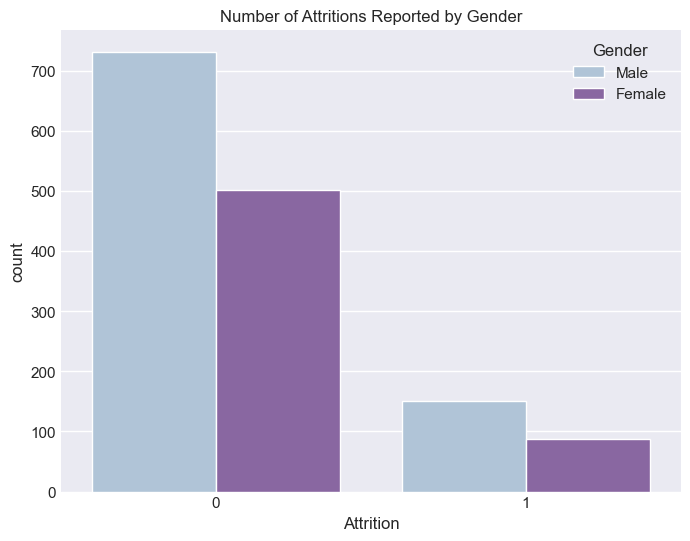
**Low representation and attrition within HR:- About 12% of workers fall into the category of Human Resources (HR), where both yes/no counts for attrition tend to be low thus indicating fewer people working there as well as less quitting from this position’s standpoint.**

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***14***

# Exploratory Data Analysis(EDA)



**Number of attritions reported by Gender :-**

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***15***



# Exploratory Data Analysis(EDA)

**Principal Takeaways from the Bar Plot:-**

**Younger Workers among Males with Higher Retention:- More than 38% of employees who did not leave (attrition = 0) are men; thus, we may conclude that young male workers demonstrate higher loyalty.**

**Medium Attrition for Both Genders:- Almost 19% of those who left were males (attrition = 1), showing that while their retention rate is high, it does not eliminate considerable loss through attrition.**

**Similarly, around 12% of leavers were women meaning that women also have modest attrition levels.**

**Females Have Significant Retention Rate:- Over 30% of employees who did not quit are females, which**

**means that a great part of the female population is retained in the company.**

**Overall Distribution:- Thus, this plot shows that most members of staff stay on while many females and males remain rather than quit.**

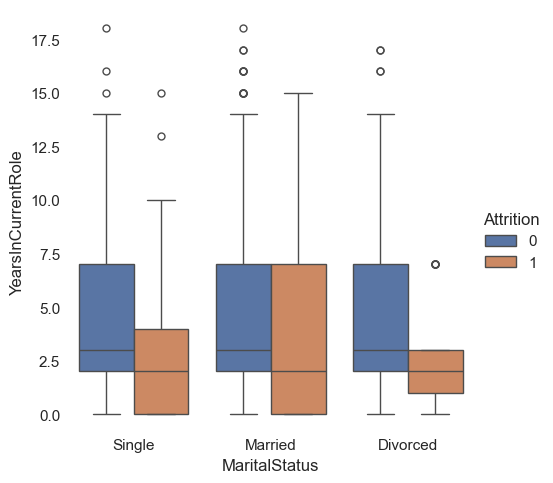
**This visualization helps to understand retention and attrition trends across gender lines thereby providing insightful thinking about stability and turnover rates within the workforce.**

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***16***

# Exploratory Data Analysis(EDA)



**Years in cureeent Role and marital status:-**



*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***17***



# Exploratory Data Analysis(EDA)

Single Individuals:- Employees who stayed (attrition = 0) had a higher median number of years in their current roles than employees who left (attrition = 1).

Those who stayed have a wider range of years in the current role than those who left.

Married People:- The median number of years in the current role for individuals that remained is also

higher than for those who left.

Both groups have a visible distribution spread with some overlap.

Divorced people:- Median years in the current role between groups, i.e. those who remained and left, look closer compared to single and married employees.

There are many outliers for both groups over significant number of years with more staying put.

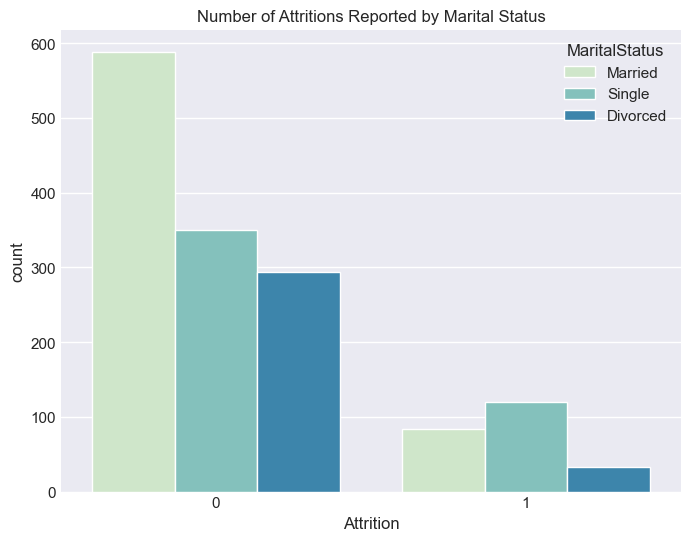
*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***18***

# Exploratory Data Analysis(EDA)



**Number of Attritions Reported by Marital Status**

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***19***



# Exploratory Data Analysis(EDA)

**Marital Status and Attrition Counts:Married Employees:-** The count of married employees who did not leave (attrition = 0) is the highest among the three groups.

The count of married employees who left (attrition = 1) is relatively low.

**S~~ingle Em~~ployees:-** The count of single employees who did not leave is lower than married employees but higher than divorced employees.

The count of single employees who left is higher than married employees but significantly higher than divorced employees.

**Divorced Employees:-**The count of divorced employees who did not leave is the lowest among the three groups.

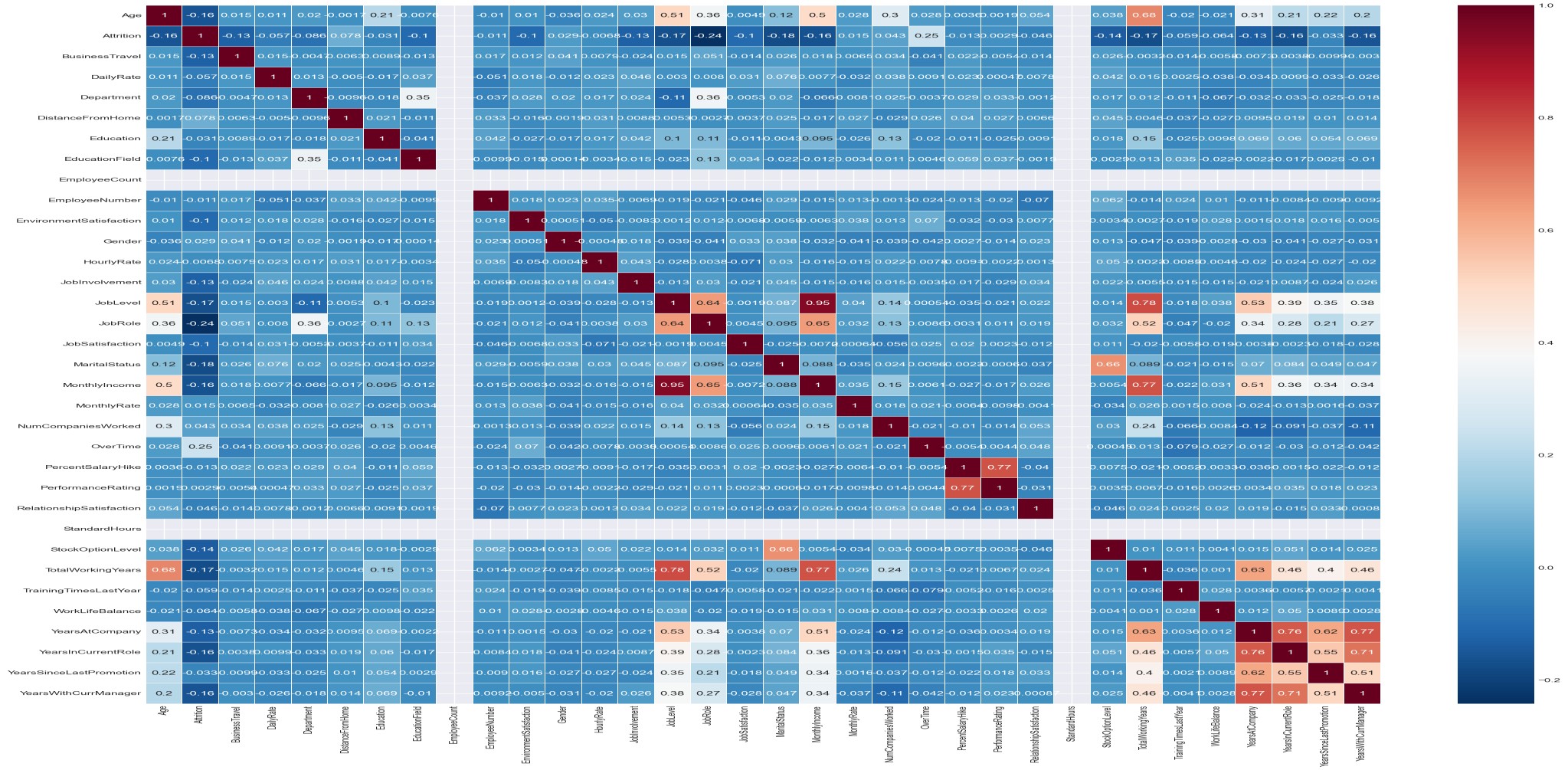
The count of divorced employees who left is also the lowest.

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***20***

# Exploratory Data Analysis(EDA)



**Number of Attritions Reported by Marital Status**



*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***21***



# Exploratory Data Analysis(EDA)

Strong Positive Correlations (Closer to 1):- Job Level and Total Working Years (0.77): The more years spent on working, the higher his job level would be. Monthly Income and Job Level (0.95): Higher job levels are strongly correlated with an increased monthly income.

Total Working Years and Monthly Income (0.78): Employees whose total

working years are higher tend to have a bigger monthly income.

Years at Company and Years with Current Manager (0.77): It is also seen that longer stay with the company implies more number of years with the same manager.

Years in Current Role and Years with Current Manager (0.70): There is an indication that employees who have stayed in their current positions for long have had their managers around for quite some time too.

Years at Company and Years in Current Role (0.76): In addition, employees who have been longest with the company also have worked for the most prolonged periods in their present role.

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***22***

# Exploratory Data Analysis(EDA)



Strong Negative Correlations (Closer to -1):-

Attrition and Age (-0.16): The likelihood of quitting is slightly lower among older employees.

Attrition and Years Since Last Promotion (-0.17): Recent

~~promo~~tions decrease the possibility of leaving the firm a bit.

Total Working Years and Attrition (-0.16): The chances of leaving the company are reduced for workers with more total working years.

Years with Current Manager and Attrition (-0.16): The chances of leaving are fewer if members have stayed long enough to learn about their current managers, than those who have not.

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***23***

# Exploratory Data Analysis(EDA)



Moderate Positive Correlations:

Job Satisfaction and Job Involvement (0.49): Job involvement increases with

an increase in job satisfaction at moderate levels.

Percent Salary Hike and Performance Rating (0.77): Increase in percentage salary hikes is associated with high performance ratings.

Moderate Negative Correlations:

Distance from Home and Job Satisfaction (-0.25): Commuting over longer

distances relates negatively to job satisfaction at moderate levels.

Work-Life Balance and Overtime (-0.24): Poorer work-life balance results

from continued overtime working at medium degree

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***24***



# Building Machine Learning Models:

In this section, we develop a supervised learning distribution model to predict employee turnover. Given that the task is to determine whether the employee will leave ("yes") or stay ("no"), we use various classification algorithms to build and improve our models. The test is set using the train\_test\_split function with a parameter size of 0.2. For our first model, we used logistic regression, a simple and effective algorithm. To ensure robustness, we tested the logistic regression model using 86 random cases and obtained an F1 score of 0.92.

Performance improvement. We used k-fold cross-validation (k = 5) and found that the random forest classifier achieved the highest F1 and average cross-validation score. This result led us to perform a hyperparameter transformation of the random forest model to improve its performance. Although the accuracy decreased slightly, the modified model still performed well. Therefore, we decided to use a random forest model with pre-tuned hyperparameters as our final model due to its strong performance and simplicity. It is integrated into the platform and production systems. This allows the model to be easily used for instant prediction.

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***25***

# Building Machine Learning Models:



After tuning, we see a slight correction from 0.87 to 0.8615. Although the accuracy drops slightly, the modified model still performs well. Therefore, we decided to use a random forest model with pre-tuned hyperparameters as our final model due to its strong performance and simplicity. It can be deployed on cloud platforms and integrated into production systems. This allows the model to be easily used for instant prediction.

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***26***



# Concluding Remarks:

Analysis of employee churn and subsequent machine learning models provide HR departments with some valuable insights and recommendations. This threshold appears to be important for employee retention because those earning less than this amount are more likely to leave.

The data reveals that attrition rates are notably high among employees aged 29 to

33. HR departments should pay close attention to the needs and expectations of this age group, as their dissatisfaction may be contributing to the elevated attrition rates.

Further analysis indicates that attrition is particularly prevalent among Sales Representatives and Laboratory Technicians. The attrition rate for Research Scientists stands at 16%, highlighting the critical nature of retaining these key employees due to their specialized skills and expertise.

Another observation is the high percentage of employees with diverse educational backgrounds in the sales department. This diversity may lead to dissatisfaction, as reflected in the high attrition rates. Addressing this issue could involve tailored training and development programs to better align employees’ expectations with organizational goals.

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***27***

# Concluding Remarks:



In terms of feature engineering, various techniques such as data analysis, outlier removal, label coding, feature selection, and fundamental analysis (PCA) are used to improve the performance of the model’s features. The random forest classifier turned out to be the best model that provided the most accuracy in predicting the change drivers. This performance highlights the importance of advanced machine learning to solve complex HR problems.

*HR Analytics Project- Machine Learning to Understand & Predict HR Attrition* ***28***