Contents

[HR ANALYTICS – WHY DO EMPLOYEES QUIT? 1](#_Toc208231487)

[Exploratory Data Analysis (EDA) 1](#_Toc208231488)

[Our Dataset 1](#_Toc208231489)

[Libraries 3](#_Toc208231490)

[Graphs 3](#_Toc208231491)

[CONCLUSION 8](#_Toc208231492)

[CORRELATION GRAPHICS 9](#_Toc208231493)

[PREDICTING EMPLOYEE ATTRITION 10](#_Toc208231494)

[GridSearchCV 10](#_Toc208231495)

[LogisticRegression 10](#_Toc208231496)

[RandomForest 10](#_Toc208231497)

[Support Vector Machine (SVM) 11](#_Toc208231498)

[XGBoost Classifier 12](#_Toc208231499)

[LightGBM Classifier 12](#_Toc208231500)

[Conclusion 13](#_Toc208231501)

# 

# HR ANALYTICS – WHY DO EMPLOYEES QUIT?

Employees are the backbone of any organization and an organization's performance is heavily based on the quality of its employees. The challenges that an organization has to face due employee attrition are:

1. Expensive in terms of both money and time to train new employees.
2. Loss of experienced employees
3. Impact in productivity
4. Impact profit

So why do employees choose to leave their companies? This question can be answered through the means of a comprehensive dataset, and EDA.

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) helps us uncover patterns, trends, and hidden insights in the data through visualizations and statistical techniques. For univariate analysis, we use histograms and boxplots to understand the distribution, central tendency, and spread of continuous variables, while bar charts help us analyse the frequency of categorical variables. In bivariate analysis, scatter plots reveal relationships between two continuous variables, stacked bar charts are useful for comparing frequencies between two categorical variables, and boxplots or swarm plots help us examine how a continuous variable varies across different categories. EDA also involves detecting outliers that may distort analysis and handling them appropriately. Finally, feature engineering plays an important role, where raw data is transformed into meaningful inputs—such as creating new features, encoding categorical values, or scaling variables—to improve the performance and accuracy of predictive models.

## Our Dataset

The dataset captures employee demographics, job-related details, compensation, satisfaction levels, and career progression within the company.

* **EmployeeCount** – Total number of employees (constant in most datasets).
* **EmployeeNumber** – Unique ID given to each employee.
* **EnvironmentSatisfaction** – Rating of satisfaction with workplace environment (scale 1–4).
* **Gender** – Employee’s gender (Male/Female).
* **HourlyRate** – Employee’s hourly wage.
* **JobInvolvement** – Measure of how involved and committed an employee is (scale 1–4).
* **JobLevel** – Employee’s level in the company hierarchy.
* **JobRole** – The specific role or designation of the employee.
* **JobSatisfaction** – Rating of satisfaction with the job itself (scale 1–4).
* **MaritalStatus** – Employee’s marital status (Single/Married/Divorced).
* **MonthlyIncome** – Employee’s monthly salary.
* **MonthlyRate** – Total monthly pay rate (different from fixed salary).
* **NumCompaniesWorked** – Number of companies the employee has worked at before.
* **Over18** – Whether the employee is over 18 years old (Yes/No).
* **OverTime** – Whether the employee works overtime (Yes/No).
* **PercentSalaryHike** – Percentage increase in salary during the last hike.
* **PerformanceRating** – Employee’s performance score (scale 1–4).
* **RelationshipSatisfaction** – Rating of satisfaction with relationships at work (scale 1–4).
* **StandardHours** – Standard working hours (usually fixed at 80).
* **StockOptionLevel** – Stock options granted to the employee (scale 0–3).
* **TotalWorkingYears** – Total years of professional experience.
* **TrainingTimesLastYear** – Number of training sessions attended in the last year.
* **WorkLifeBalance** – Rating of balance between work and personal life (scale 1–4).
* **YearsAtCompany** – Number of years the employee has worked at the company.
* **YearsInCurrentRole** – Number of years spent in the current role.
* **YearsSinceLastPromotion** – Number of years since the employee’s last promotion.
* **YearsWithCurrManager** – Number of years the employee has worked under their current manager.

Evidently, the dataset is highly informational, comprising of various columns elucidating multiple aspects of the employee’s work life; thereby making it perfectly suitable for a detailed analysis on employee attrition.

## Libraries

We’ll be using **hvPlot** to make our charts.

hvPlot is great because it works directly with data in a Pandas DataFrame, so you can make nice, interactive plots with just one line of code. It’s simple, fast, and really useful when you want to quickly explore patterns in your data.

For dataset analysis, Pandas is the library of choice.

## Graphs

A graph of a number of years

AI-generated content may be incorrect.A graph of a number of years

AI-generated content may be incorrect.A graph of blue and red bars

AI-generated content may be incorrect.A graph of a number of people

AI-generated content may be incorrect.A graph of a number of people

AI-generated content may be incorrect.A graph with blue and red bars

AI-generated content may be incorrect.A graph of a bar graph

AI-generated content may be incorrect.A graph of different colored columns

AI-generated content may be incorrect.A graph with blue and red bars

AI-generated content may be incorrect.A graph of a bar graph

AI-generated content may be incorrect.A graph of different colored bars

AI-generated content may be incorrect.A graph of a bar chart

AI-generated content may be incorrect.A graph of a graph with a number of columns

AI-generated content may be incorrect.A graph of a person with a blue and red line

AI-generated content may be incorrect.

A graph of blue and red squares

AI-generated content may be incorrect.

## CONCLUSION

* Employees with **lower job levels, lower monthly income, fewer years at the company, and fewer total working years** are more likely to leave.
* Those who **travel frequently for work** are also more likely to quit compared to employees who travel less.
* People in the **Research & Development department** tend to stay longer than those in other departments.
* Workers with **Human Resources or Technical degrees** are more likely to quit than employees from other educational fields.
* **Men** are more likely to leave compared to women.
* Employees working as **Laboratory Technicians, Sales Representatives, or in Human Resources roles** have higher chances of leaving compared to other job roles.
* Workers who are **single** are more likely to quit than those who are married or divorced.
* Employees who **regularly work overtime** are more likely to quit than those who don’t.

## CORRELATION GRAPHICS

A screenshot of a computer screen

AI-generated content may be incorrect.

A graph with blue rectangles

AI-generated content may be incorrect.

* Monthly income is highly correlated with Job level.
* Job level is highly correlated with total working hours.
* Monthly income is highly correlated with total working hours.
* Age is also positively correlated with the Total working hours.
* Marital status and stock option level are negatively correlated

## PREDICTING EMPLOYEE ATTRITION

In order to predict whether a certain employee will leave the company or stay, we use Python’s Scikit-Learn library and various models to determine which method can provide the highest accuracy.

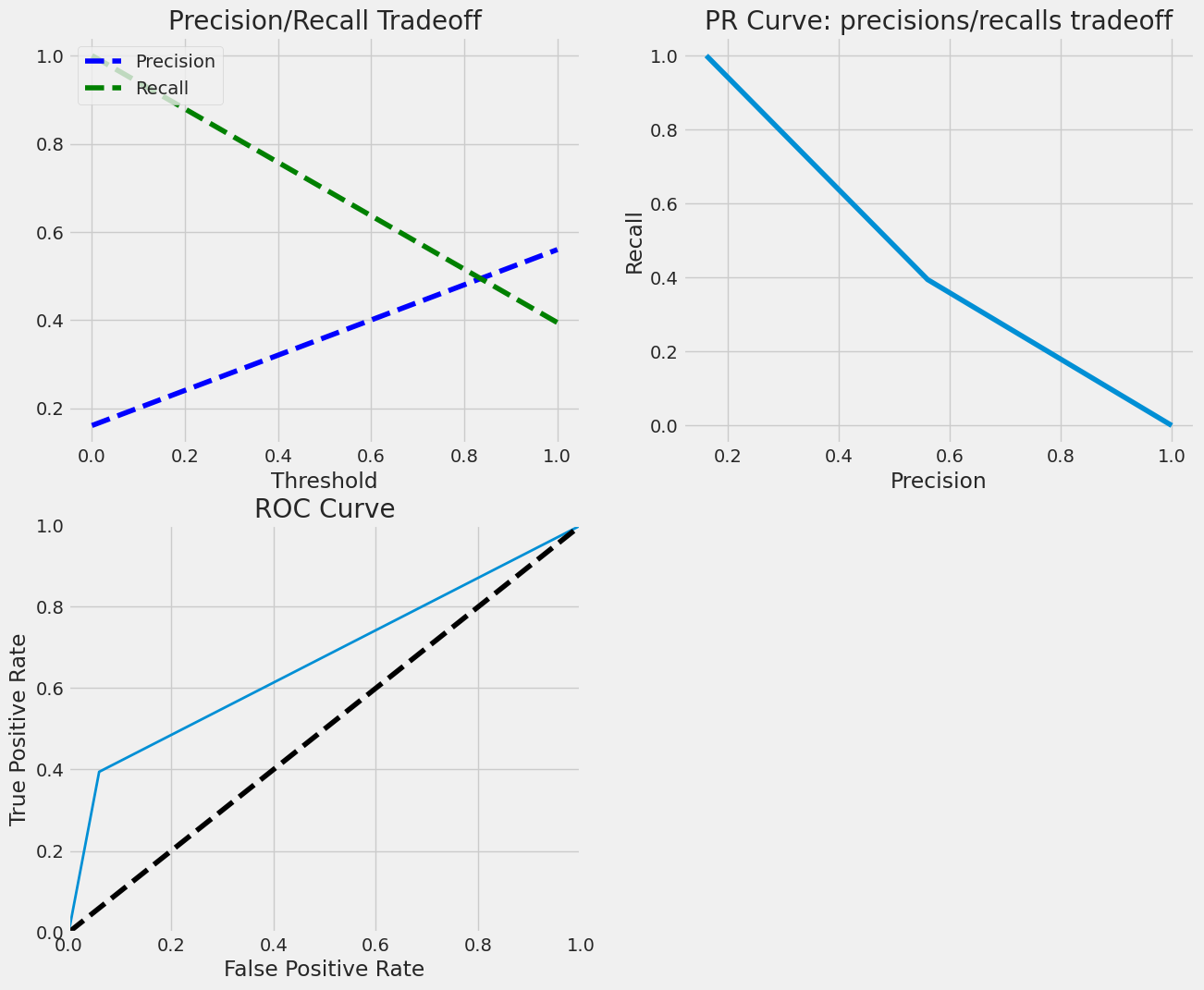
### GridSearchCV

We proceed by splitting the dataset into a training and test set, before applying GridSearchCV on both components.

**Seeing that our dataset is unbalanced, the model manages to predict employee attrition with an 83.9% accuracy.**

### LogisticRegression

Upon applying Logistic regression to train and test set, we receive an accuracy of 85% in test set.



### RandomForest

Using Random Forest Classifier, we receive an accuracy of 84.3% in testing set with MonthyIncome being given the highest importance.

A graph of a graph of a graph

AI-generated content may be incorrect.

### Support Vector Machine (SVM)

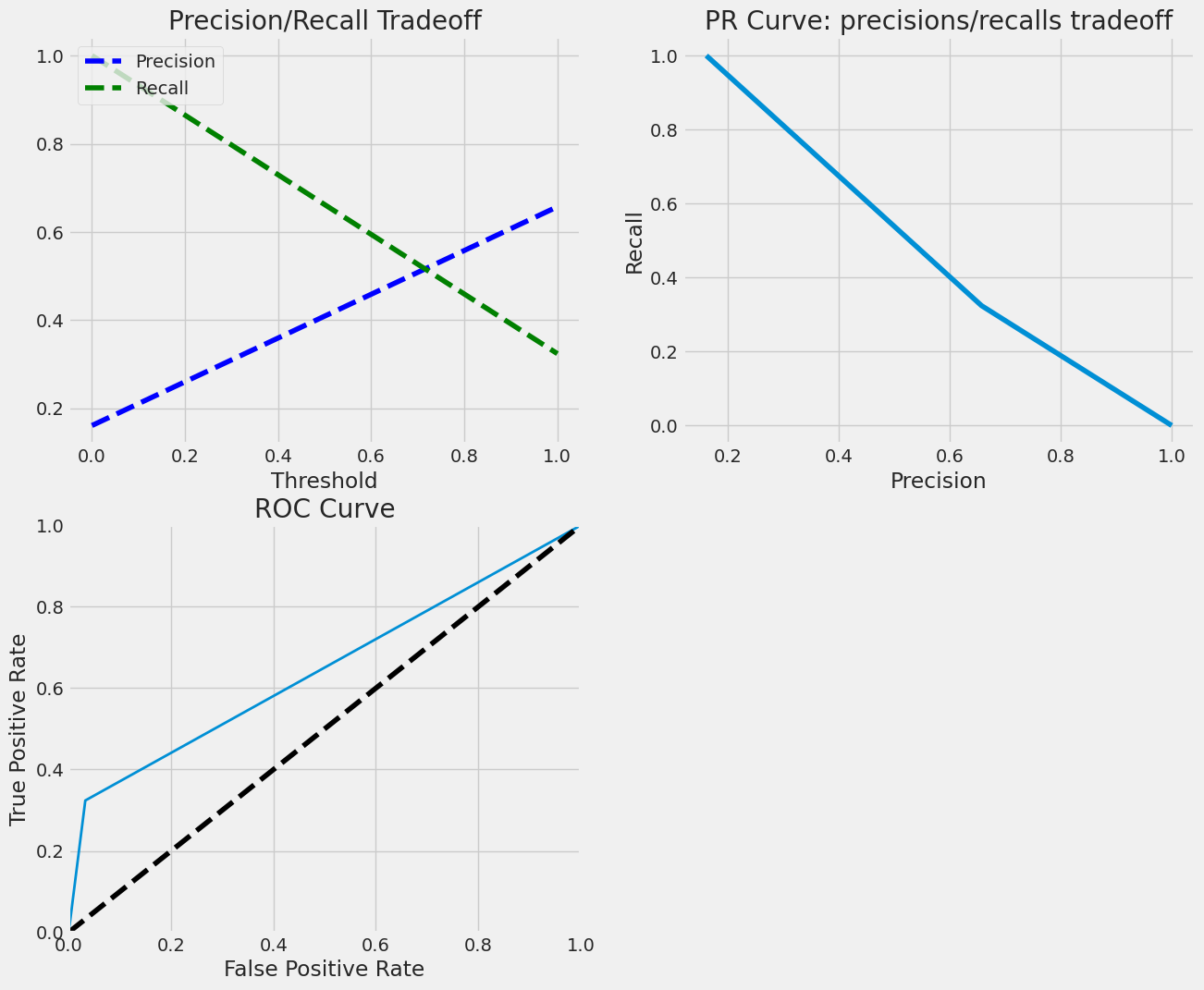
Using a SVM model, we receive an accuracy of 85.0% in the test set.

A graph of different types of graphs

AI-generated content may be incorrect.

### XGBoost Classifier

Using XGBoost classifier, we receive an accuracy of 86.3% in the test set.



### LightGBM Classifier

The LightGBM Classifier yields an accuracy of 85.0% in the test set.

A graph of different types of graphs

AI-generated content may be incorrect.

### Conclusion

Across different machine learning models, accuracy scores for predicting employee attrition ranged between **83.9% and 86.3%**. Among them, **XGBoost performed the best** with an accuracy of **86.3%**, closely followed by Logistic Regression, SVM, and LightGBM at around **85%**. Random Forest also provided strong results with **84.3%**, while highlighting **Monthly Income** as the most influential factor in predicting attrition.

Overall, these results show that advanced ensemble methods like **XGBoost** and **LightGBM** slightly outperform traditional models, but even simpler models such as **Logistic Regression** remain highly competitive and interpretable for this task.