Predicting Employee Attrition Using Classification Model

Riti Dahal

University of South Dakota

DSCI 519 Advanced Business Analytics Modeling

Dr. Hanus

December 10, 2025

Table of Contents

Introduction

Employee attrition is a major problem as losing skilled employees results in higher hiring expenses, decreased productivity, and a loss of organizational competence. The main objective of this project is to use the IBM HR Analytics Employee Attrition dataset to develop predictive classification models. This model will help to identify important factors influencing employee turnover. This project uses Python-based machine learning algorithms to explore relationships across demographic, job-related, and performance variables. The project further analyzes the degree to which these factors impact an employee's probability of leaving the organization.

HR professionals and business leaders can gain practical insights by comprehending these factors that contribute to turnover. Predictive analytics can help HR professionals identify high-risk employee categories, enhance workforce planning, and support proactive retention strategies. These insights can help organizations make better decisions about professional development opportunities, compensation practices, and employee engagement. This research shows how analytics may be used to assist evidence-based HR management and reduce avoidable turnover by combining structured data preparation, exploratory analysis, and model comparison.

Data Exploration

The dataset used for this analysis was sourced from Kaggle which provides information about IBM Employee Attrition. After importing the dataset into Google Collab, it shows that the dataset contains 1470 employee records and 35 variables including demographics, job characteristics, compensation, performance and work-life factors. Out of the 35 variables, 26 are numerical variables and 9 are categorical variables. Further, it was found that the data set has no missing value which simplifies the data preparation process.

Both the numerical and categorical variables are listed in the table below:

Numerical Variables

Variable Name	Description
Age	Employee age
DailyRate	Daily salary rate
DistanceFromHome	Distance between home and workplace (miles)
Education	Education level (1–5)
EmployeeCount	Employee count indicator (constant = 1)

EmployeeNumber	Unique employee ID
EnvironmentSatisfaction	Work environment satisfaction (1–4)
HourlyRate	Hourly salary rate
JobInvolvement	Job involvement level (1–4)
JobLevel	Job seniority level
JobSatisfaction	Job satisfaction (1–4)
MonthlyIncome	Monthly salary
MonthlyRate	Monthly pay rate
NumCompaniesWorked	Number of companies previously worked at
PercentSalaryHike	Percent salary increase
PerformanceRating	Performance rating (1–4)
RelationshipSatisfaction	Relationship satisfaction (1–4)
StandardHours	Standard working hours per week (constant = 80)
StockOptionLevel	Stock option level (0–3)
TotalWorkingYears	Total years of professional experience
TrainingTimesLastYear	Number of trainings completed last year
WorkLifeBalance	Work–life balance score (1–4)
YearsAtCompany	Years at current company
YearsInCurrentRole	Years spent in current job role
YearsSinceLastPromotion	Years since last promotion
YearsWithCurrManager	Years with current manager

Categorical Variables

Variable Name	Description
Attrition	Indicates whether the employee left the organization (“Yes” or “No”). This is the target variable.
BusinessTravel	Frequency of business travel (e.g., Non-Travel, Travel_Rarely, Travel_Frequently).
Department	Department where the employee works (e.g., Sales, R&D, HR).
EducationField	Field of study (e.g., Life Sciences, Medical, Marketing).
Gender	Employee gender.
JobRole	Specific job role or position (e.g., Sales Executive, Laboratory Technician).
MaritalStatus	Marital status (e.g., Single, Married, Divorced).
OverTime	Indicates whether the employee works overtime (Yes/No).
Over18	Indicates whether the employee is over 18 (constant value in this dataset).

One-hot encoding was subsequently used to encode these categorical features in order to guarantee machine-learning model compatibility.

Summary Statistics for Key Numerical Variables

The tables below present the key numerical variables that are most relevant for understanding attrition patterns. These metrics provide valuable insights into employee demographics, compensation levels, satisfaction scores, and tenure within the organization. Because the dataset contains a large number of numerical variables, the summary statistics have been organized into two separate tables for improved readability and clarity.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Age	1470	36.92	9.14	18	30	36	43	60
DailyRate	1470	802.49	403.51	102	465	802	1157	1499
DistanceFromHome	1470	9.19	8.11	1	2	7	14	29
Education	1470	2.91	1.02	1	2	3	4	5
EmployeeCount	1470	1.00	0.00	1	1	1	1	1
EmployeeNumber	1470	1024.87	602.02	1	491	1020.5	1555.75	2068
EnvironmentSatisfaction	1470	2.72	1.09	1	2	3	4	4
HourlyRate	1470	65.89	20.33	30	48	66	83.75	100
JobInvolvement	1470	2.73	0.71	1	2	3	3	4
JobSatisfaction	1470	2.73	1.10	1	2	3	4	4
MonthlyIncome	1470	6502.93	4707.96	1009	2911	4919	8379	19999
Variable	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**

MonthlyRate	1470	14313.10	7117.79	2094	8047	14236	20461.5	26999
NumCompaniesWorked	1470	2.69	2.50	0	1	2	4	9
PerformanceRating	1470	3.15	0.36	3	3	3	3	4
RelationshipSatisfaction	1470	2.71	1.08	1	2	3	4	4
StandardHours	1470	80.00	0.00	80	80	80	80	80
StockOptionLevel	1470	0.79	0.85	0	0	1	1	3
TrainingTimesLastYear	1470	2.80	1.29	0	2	3	3	6
WorkLifeBalance	1470	2.76	0.71	1	2	3	3	4
YearsAtCompany	1470	7.01	6.13	0	3	5	9	40
YearsSinceLastPromotion	1470	2.19	3.22	0	0	1	3	15

Interpretation of Summary Statistics

- The average employee is 37 years old, with ages ranging from 18 to 60.
- Monthly income varies widely, from $1,009 to $19,999, with an average of $6,503.
- Employees live an average of 9 miles from work, with commutes ranging from 1 to 29 miles.
- Employees have worked for their company for an average of 7 years, but tenure ranges from 0 to 40 years.
- The number of companies employees previously worked for ranges from 0 to 9, with an average of 2.7.
- Time since last promotion varies greatly, from 0 to 15 years, with a median of 1 year.
- Satisfaction scores (environment, job, relationship) average around 2.7–2.8 on a 1–4 scale, showing generally moderate satisfaction.
- Training sessions last year ranged from 0 to 6, with employees completing an average of 2–3 sessions.
- Some variables (e.g., Employee Count and Standard Hours) remain constant and therefore do not provide predictive value.

Attrition Distribution

In this analysis, the main focus is attrition, which shows whether an employee stays or leaves the company. It is recorded as a binary variable: “Yes” for employees who left and “No” for those who remained. Since employee turnover can significantly affect workforce stability, productivity, and organizational costs, it is imperative to understand attrition patterns. Thus, by analyzing the factors associated with attrition, HR teams can identify which employees are more likely to leave and based on that HR can further develop retention strategies.

Attrition	**Percentage (%)**
No	83.88%
Yes	16.12%

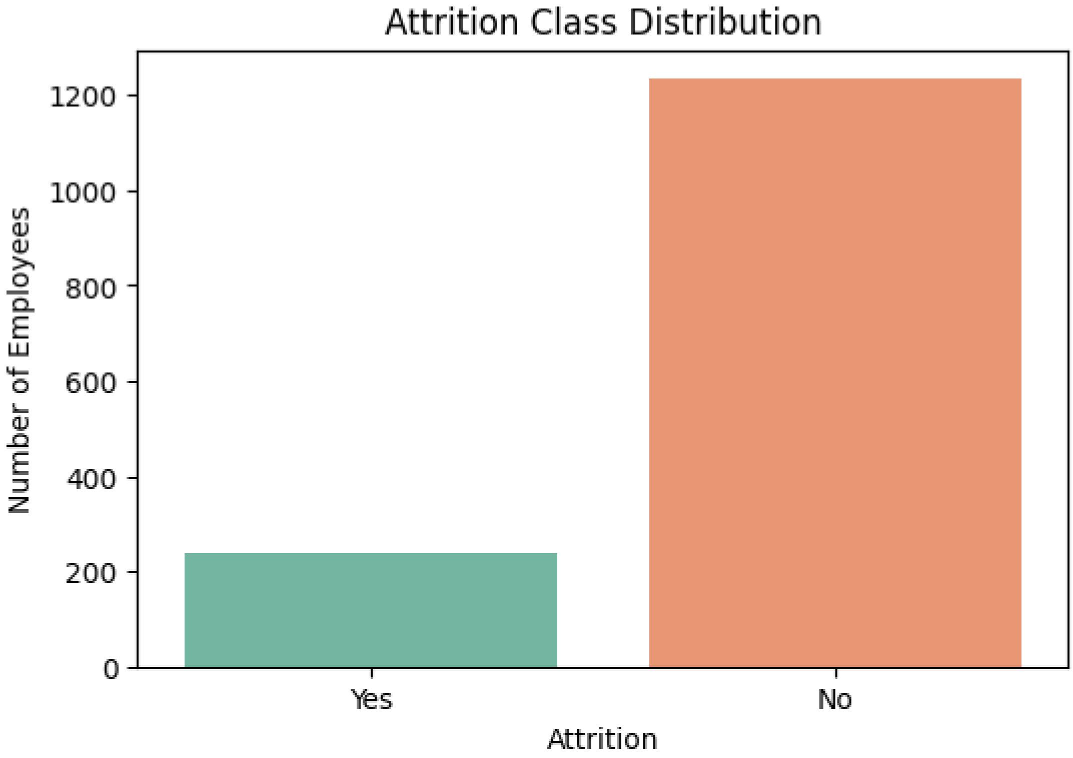

Fig 1: Attrition Class Distribution

The attrition distribution indicates that a large percentage of employees in the dataset stayed with the company. In this case, only 237 employees (16.12%) left the company, while 1233 employees (83.88%) stayed. With attrition cases making up a far smaller percentage of the workforce, this suggests that the data set is highly imbalanced.

Attrition by Over Time

The bar plot shows a strong relationship between overtime and attrition. Employees who work overtime have a noticeably higher number of attrition cases compared to those who do not work overtime. While the majority of the workforce does not work overtime, the attrition count among overtime workers is significantly higher. This shows that overtime workload contributes to employee burnout, hence increasing the likelihood of an employee leaving the company.

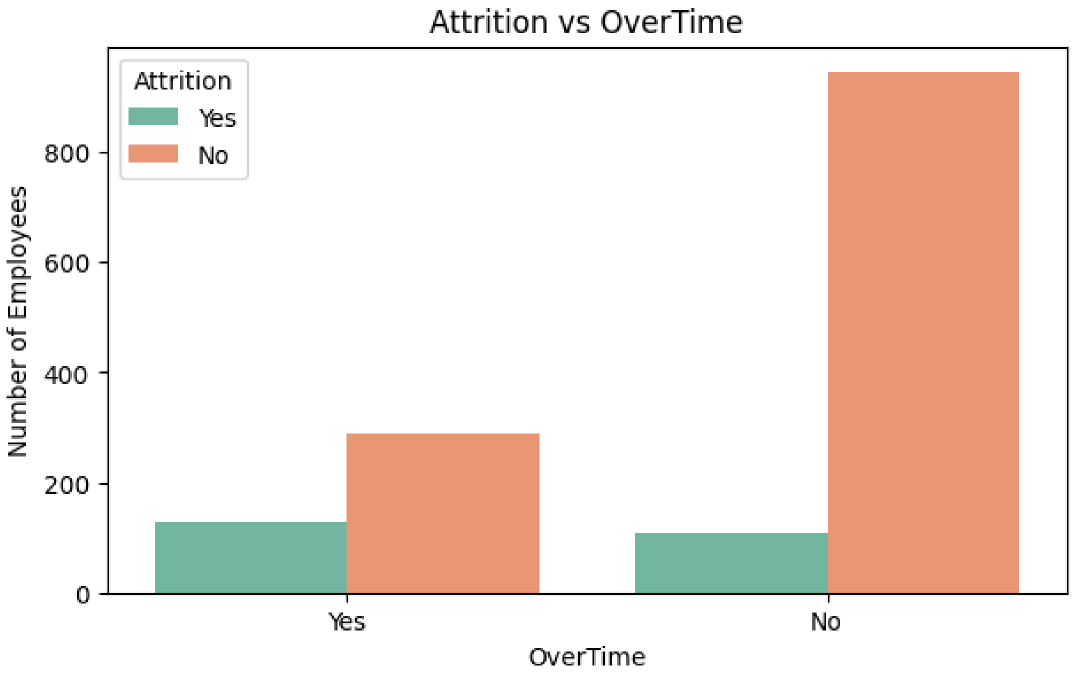

Fig 2 : Attrition by Over Time

Attrition by Job Role

The plot shows how attrition rates differ among various employment roles. When compared to positions like Manager, Research Director, or Human Resources, which have extremely low attrition, positions like Laboratory Technician, Sales Executive, and Research Scientist show a higher employee turnover. These differences show that some job roles with demanding workloads, limited growth, or low salaries are more vulnerable to turnover. Therefore, this project can help an organization focus on its retention strategies and help them understand which roles have higher attrition rates.

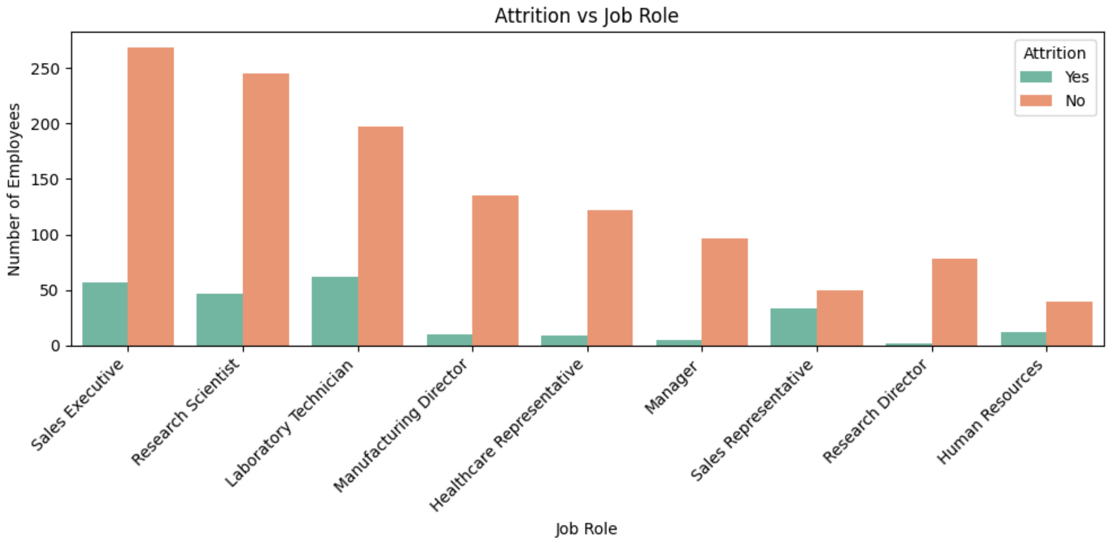

Fig 3: Attrition by Job Role

Attrition by Monthly Income

The bar plot shows the attrition by monthly income. As we can see, employees who left the company have lower monthly incomes compared to those who stayed. The median income for the attrition group is much lower, while employees who remain have a wider range of higher salaries. This shows that lower compensation may be linked to a higher likelihood of leaving the company.

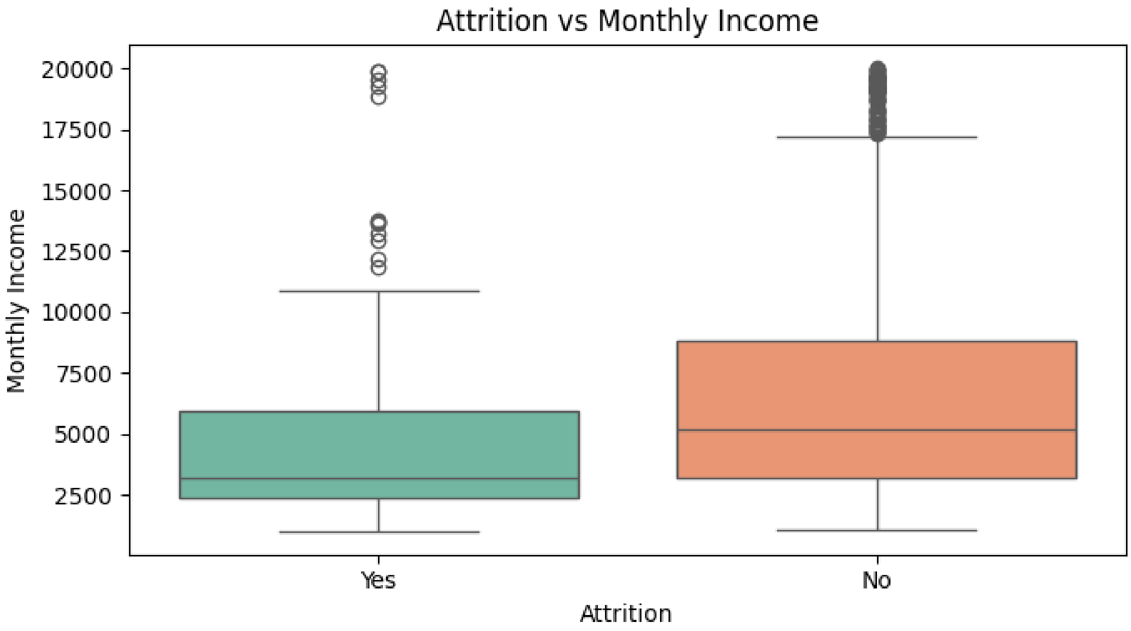

Fig 4: Attrition by Monthly Income

Summary of Age by Attrition

The plot shows that attrition is more common among younger employees, especially those in their 20s and early 30s. As age increases, the number of employees leaving the company decreases significantly. This suggests that younger workers may be more likely to switch jobs or leave for better opportunities.

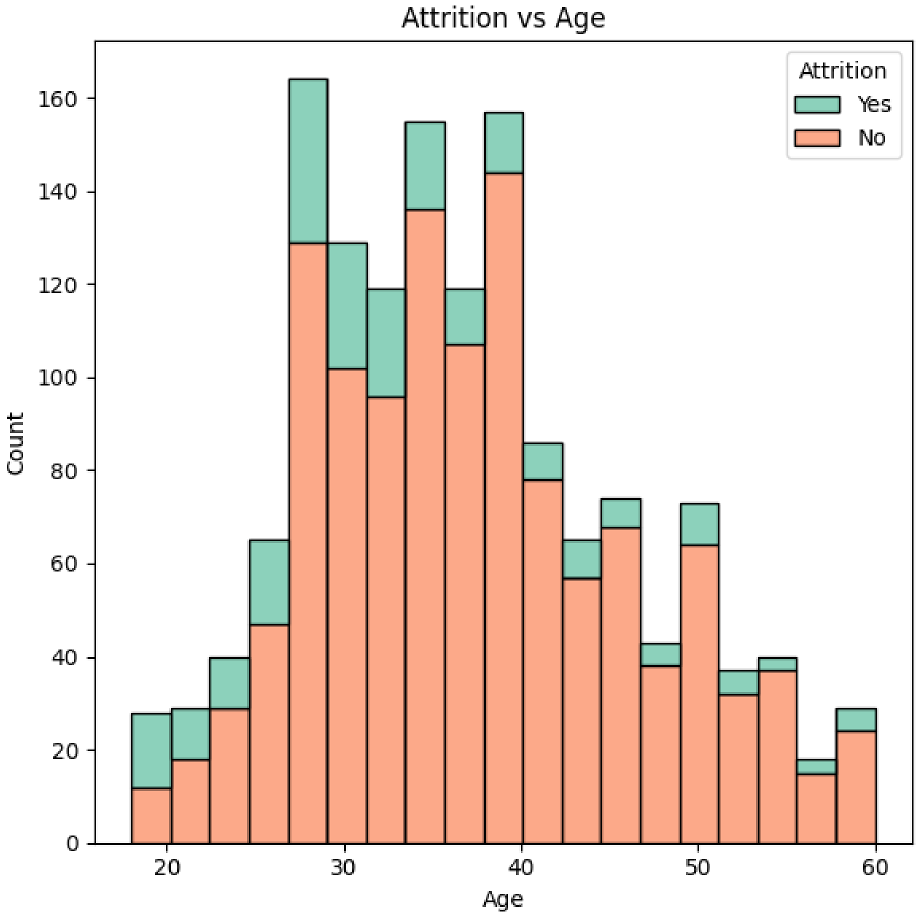

Fig 5: Summary of Age by Attrition

Correlation Analysis

The correlation heatmap shows how numerical variables in the dataset are related to each other. From the plot we can see Monthly Income and Job Level have a very high correlation of (0.95), and Total Working Years and Monthly Income correlate at (0.78), showing that income, experience, and job level rise together. Tenure-related variables also move together, such as Years at Company and Years with Curr Manager (0.77). In contrast, satisfaction scores such as

Job Satisfaction and Environment Satisfaction show very weak correlations with most variables which are between (–0.07 and 0.07), meaning they are mostly independent. Distance from home, training counts, and rate variables also show weak relationships (close to 0.00–0.10). Overall, the heatmap indicates strong relationships among career progression measures, while many other employee characteristics are relatively unrelated.

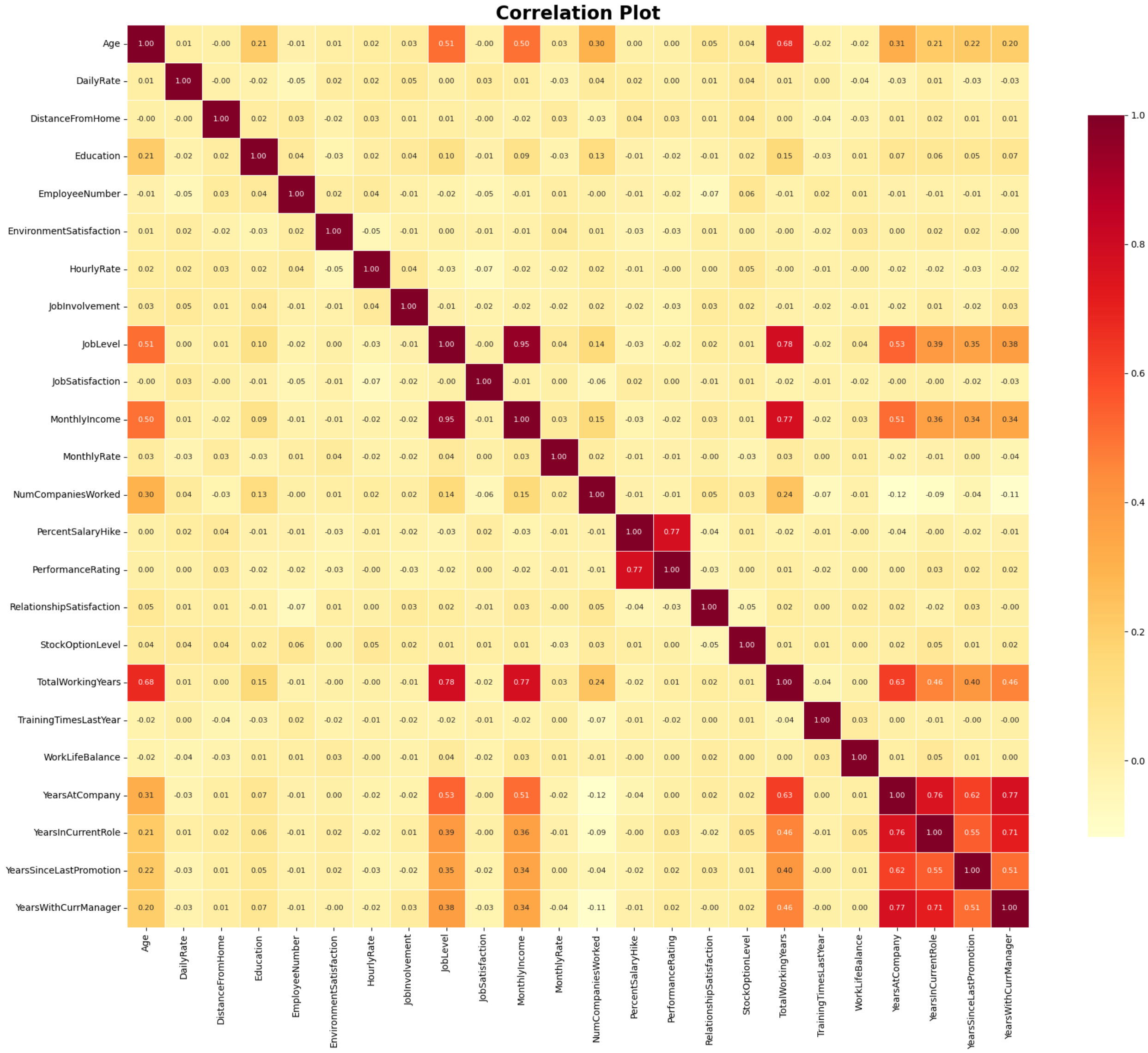

Fig 6: Correlation Analysis

Data Transformation and Data Preparation

After looking at the correlation plot, the five variables including Job Level, Total Working Years, Percentage Salary Hike, Years in Current Role and Years with Current Manager were removed as they were highly correlated with other features in the dataset. Dropping these variables helps to reduce multicollinearity, avoids redundancy, and improves model interpretability and stability, especially when building a model like Logistic Regression.

Similarly, the Employee Number column in the dataset was also removed as it is an identifier and does not provide meaningful information for prediction. Further, to prepare the data for classification, the target variable Attrition was converted into a binary format. Employees who left the company were assigned as 1 and who stayed in the company were assigned as 0. The remaining columns were separated into numerical and categorical variables and used as predictors. Also, the categorical data were converted using one-hot encoding, which turns each category into a distinct binary column.

In order to prevent multicollinearity and the creation of redundant dummy variables, the parameter drop_first=True was used. Following encoding, the columns were ordered alphabetically to ensure consistency across training and testing sets, and the number of features was expanded to reflect the entire category of complexity of the dataset. Lastly, the standard scaler was used to normalize the numerical variables to adjust each feature to have a mean of 0 and a standard deviation of 1. This process helps to prevent the large range of variables from dominating the model and improves coefficient stability and interpretability.

Data Partitioning

The dataset was split into training and testing sets to support reliable model development and evaluation. A 70/30 split was applied using the train_test_split function in Python, where 70 % of the data was used to train the models and learn the underlying patterns in employee attrition. The remaining 30% of the data was used for the testing set to assess how well the model performs on unseen data.

Likewise, stratified sampling was used to ensure that the proportion of employees who stayed and those who left remained consistent in both sets, which is important due to the imbalance in the Attrition variable. During the splitting procedure, a fixed random seed of 42 was used to provide reliable results and the ability to generate the same partition again if necessary. There were 1,029 observations in the final training set and 441 observations with 41 features each in the testing set.

Dataset	**Number of Observations**	**Number of Features**
Training Set	1029	41
Testing Set	441	41

Table 1: Data Partitioning

Model Evaluation Techniques

The models were evaluated using several key performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. In addition, a confusion matrix was generated for each model to show how well it classified employees who stayed versus those who left. These evaluation techniques provided a balanced understanding of overall performance, especially for the minority attrition class.

Handling Class Imbalance Using SMOTE

SMOTE was implemented to balance the training data as there were significantly less employees in the dataset who departed the organization than those who stayed. By producing artificial samples of the minority class, SMOTE gives the models equal representation of both groups and improves their ability to learn attrition trends. SMOTE was used for both unscaled data for tree-based models and scaled data for models such as Logistic Regression, SVM, and Neural Networks. Both classes had the same number of data after balancing, which enhanced the model's capacity to identify employees who were at risk of leaving.

Class (Attrition)	Before SMOTE	After SMOTE
0 (Stayed)	863	863
1 (Left)	166	863

Table 2: Class Imbalance

Baseline Model: Logistic Regression

Logistic Regression was used as the baseline model for predicting employee attrition. The table below shows the top coefficients from the model which affects the probability of leaving the company.

Feature	Coefficient
JobRole_Sales Executive	1.139229
JobRole_Laboratory Technician	0.885372
YearsAtCompany	-0.852548
OverTime_Yes	0.845817
JobRole_Sales Representative	0.814728

Table 3: Top 5 Coefficient

From the above table, we can see that Job Roles like Sales Executive (1.139) and Sales Representative (0.814) both have strong positive coefficients. This shows that employees in sales roles are likely to leave the company. This might be due to high performance pressure, competitive job markets, or incentive driven environment in sales, which may encourage employees to switch jobs more frequently. Likewise, Lab Technician (0.885) also shows a positive relationship with attrition. This indicates potential challenges with limited career paths or repetitive work that may influence employees to leave the job. Another significant predictor is Over Time (0.846), which indicates that workers who put in extra hours are more likely to feel stressed or overwhelmed, which increases the possibility that they would leave. However, employees who work for years in a company, has a significant negative coefficient (-0.853). This shows that employees with longer tenure are less likely to quit, supporting the idea that employees tend to become more secure and engaged the longer they work for the same company.

Metric	Value
Accuracy	0.791383
Precision	0.411765
Recall	0.690141
F1-Score	0.515789
ROC-AUC	0.803159

Table 4: Model Performance

From the model performance table, we can see that the Logistic Regression model performed well as a baseline model. With an accuracy of (0.79), the algorithm accurately identified roughly 79% of employees as either staying or leaving. A strong ability to differentiate between employees who stay and those who leave is indicated by the ROC-AUC of (0.80). The model is good at detecting employees who are at risk of attrition, as demonstrated by the recall score of (0.69), which is particularly helpful to develop intervention strategies.

Model Performance Summary across All Models

The performance of five machine learning models was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Overall, the results show meaningful differences in how each model identifies employee attrition.

Table 5: Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Gradient Boosting	0.825397	0.442308	0.323944	0.373984	0.761591
Neural Network (MLP)	0.820862	0.437500	0.394366	0.414815	0.749524
SVM (RBF)	0.818594	0.428571	0.380282	0.402985	0.734431
Random Forest	0.834467	0.477273	0.295775	0.365217	0.722383
Decision Tree	0.739229	0.310345	0.507042	0.385027	0.637400

Decision Tree

The Decision Tree model performed the weakest among all models. It had an accuracy of (0.74) and a ROC-AUC of (0.64), which shows the limited ability to distinguish employees who stay from those who leave. Its recall was relatively higher at (0.51), identifying more attrition cases than some other models, but its low precision (0.31) shows it also produced many false positives. Overall, the Decision Tree was less accurate and less reliable compared to the more advanced models.

Random Forest

The Random Forest model achieved the highest accuracy at (0.83), but its recall was low at (0.30). This demonstrates that it missed many employees who were actually likely to leave. Its ROC-AUC of (0.72) shows moderate ability to distinguish the two classes, but the low recall suggests the model mostly predicted employees as staying. Therefore, Random Forest might not be particularly helpful for HR managers who are looking to identify potential attrition cases.

Gradient Boosting

With the highest ROC-AUC of (0.76), Gradient Boosting demonstrated the best overall performance, making it the most effective at differentiating between employees who remain and those who depart. Additionally, it achieved a strong accuracy of (0.83), but its recall of (0.32) suggests that some attrition cases were still overlooked. Gradient Boosting is one of the best models in the analysis overall due to its high predictive precision and great accuracy.

Support Vector Machine (SVM)

The SVM model performed consistently, with an accuracy of 0.82 and a ROC-AUC of 0.73. Its recall of 0.38 shows it identified a fair number of employees at risk of leaving. While not as strong as Gradient Boosting, SVM provided a balanced and reliable performance for predicting attrition.

Neural Network Model

With an accuracy of (0.82), recall of (0.39), and ROC-AUC of (0.75), the Neural Network model demonstrated balanced performance. Overall, the MLP was a good predictive model for this dataset since it was able to identify a significant number of attrition cases while still having adequate classification ability.

Feature Importance Analysis (Tree-Based Models)

The most crucial features for predicting employee attrition were determined using the tuned Decision Tree, Random Forest, and Gradient Boosting models. As we can see, Stock Option

Level and Monthly Income consistently showed up as the best predictors in all three models. Higher compensation and stock option levels are typically linked to decreased attrition risk, indicating that more financial incentives may aid in employee retention.

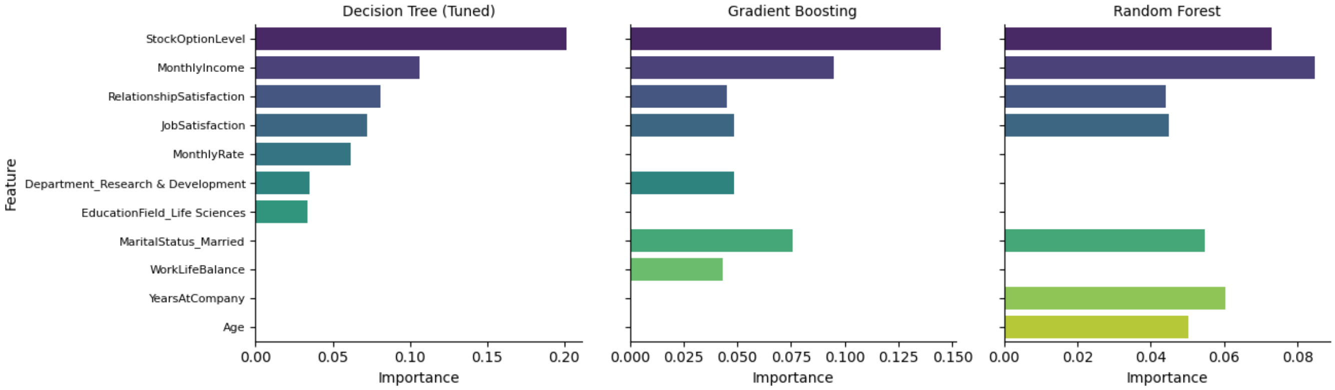

Fig 7: Feature Importance

Further, the models highlighted the importance of satisfaction related factors like job satisfaction and relationship satisfaction. This suggests that employees who express dissatisfaction with their work or relationships at work are more likely to leave the company. Additionally, variables such as Monthly Rate and Marital Status demonstrated significant impact, highlighting that an employee's decision to stay or go might also be influenced by personal circumstances and compensation structure. Overall, the three-based models indicate that this organization's attrition is driven by a combination of personal factors, satisfaction, and compensation.

Hyperparameter Tuning and Tuned Model Performance

Using GridSearchCV with 5-fold cross-validation and ROC-AUC as the scoring metric, hyperparameter tweaking was performed to enhance each model's performance by identifying the optimal set of parameters. Models can be tuned to improve their generalization, prevent overfitting, and increase predictive accuracy.

Table 6: Tuned Model Performance

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.773243	0.373913	0.605634	0.462366	0.797488
Neural Network (MLP)	0.836735	0.492537	0.464789	0.478261	0.767872
Gradient Boosting	0.848073	0.566667	0.239437	0.336634	0.736049
Random Forest	0.834467	0.477273	0.295775	0.365217	0.722383
SVM	0.836735	0.488372	0.295775	0.368421	0.710316
Decision Tree	0.721088	0.275862	0.450704	0.342246	0.628093

From the tuned model performance table, we can see that, with a ROC-AUC of (0.80) and a recall of (0.61), Logistic Regression continued to be one of the best models after tuning, and indicating that it accurately predicted a large number of employees who were likely to leave. With an accuracy of (0.84) and ROC-AUC of (0.77), the Neural Network (MLP) again demonstrated strong performance, providing a decent balance across measures. Gradient Boosting missed more attrition cases, as shown by its poor recall (0.24) despite achieving the maximum accuracy (0.85). With ROC-AUC values of roughly (0.72) and (0.71), Random Forest and SVM performed moderately, but the Decision Tree continued to be the poorest model with the lowest ROC-AUC (0.63).

ROC Curve Analysis

The ROC Curve shows the predictive performance of the top three models. With an AUC of 0.80, logistic regression displays the strongest curve, suggesting the highest capacity to differentiate between employees who stay and those who leave. With an AUC of 0.74 and 0.71, respectively, Gradient Boosting and SVM outperform a random classifier. The ROC curve provides additional evidence that Logistic Regression is the most reliable model.

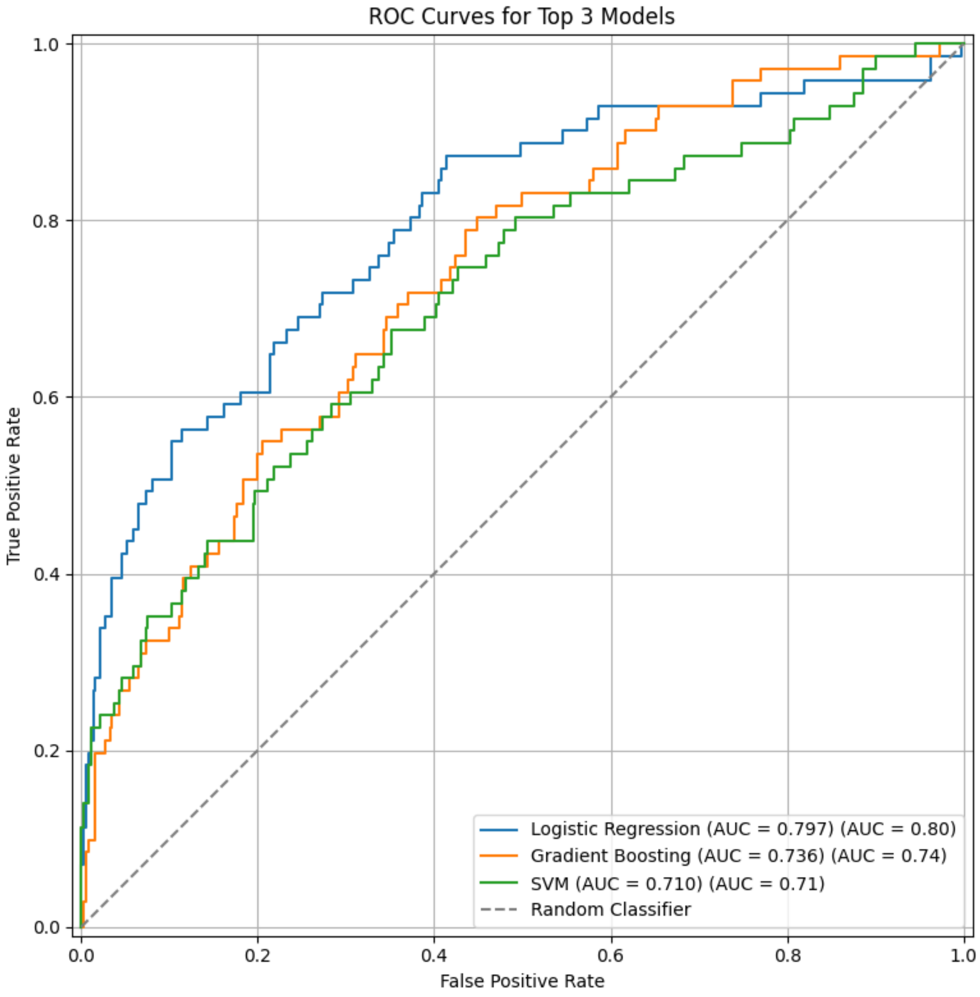

Fig 8: ROC Curve

HR Recommendations

The model's results provide several practical recommendations for reducing employee attrition. HR should give top priority to providing support for high-risk jobs, such as sales and laboratory technician professions, where turnover is more frequent. It is also important to reduce excessive overtime since those employees who work longer hours are more likely to quit. Job and relationship satisfaction can be increased by frequent check-ins, feedback surveys, and team-building exercises, helping the early resolution of workplace issues. Moreover, since financial considerations in this model were significant predictors, analyzing compensation and stock option advantages would probably yield better retention. Lastly, the HR could identify those workers who might be at risk of leaving and help them proactively through an early-warning system using the Logistic Regression model.