Analysis of Employee Attributes for Attrition Prediction and Classification

BAŞAK KABALOĞLU
FİTNAT KOÇ
SABAHATTİN ALP KOCABAŞ

***Abstract*—This research focuses on employee attrition rates based on a human resources database containing demographic characteristics and attributes defining work-life balance. The main objective is to discover which attributes are linked to employees' tendency to leave the company and to make a predictive classification.An in-depth exploratory data analysis is pursued to look for attributes' distribution, association, and potential anomalies. In this research, the original database lacks missing values; therefore, artificial missingness is introduced with a Missing-Completely-at-Random (MCAR) missingness mechanism.**

***Statistical classification models, such as logistic regression and regularized models, are used after the execution of feature engineering tasks and dimension reduction through PCA. Performance evaluation of these models is carried out using a variety of metrics such as accuracy, sensitivity, specificity, F1, and ROC-AUC, with a focus on issues pertaining to class imbalance. Key findings on the factors that contribute to employee attrition, with the use of statistical models, are inferred from the results.**

***Keywords*—Employee Attrition, Exploratory Data Analysis, Missing Data, Imputation, Logistic Regression, PCA**

INTRODUCTION

Employee attrition is a significant problem for companies as it negatively impacts productivity and increases operating costs. Therefore, identifying the factors related to employee absenteeism is crucial for effective human resources management. In this study, an employee dataset containing demographic, job-related, and work-life balance variables was analyzed to understand patterns related to turnover. The analysis involved exploratory data analysis (EDA), handling missing data, and applying predictive modeling techniques to classify employees according to their likelihood of leaving the organization. By combining EDA with predictive modeling, this study aims to provide data-driven insights that can support informed employee retention strategies and human resources policies.

LITERATURE REVIEW

Employee attrition has been examined in the literature from multiple perspectives. Some studies focus on analyzing employee behavior to identify the underlying factors that influence employees' decisions to remain in or leave an organization [1]. Other studies approach the problem from a predictive perspective by applying machine learning techniques to human resource datasets. Multiple classification models such as random forests, k-nearest neighbors, and support vector machines were evaluated using different versions of the IBM HR dataset, including the original imbalanced dataset as well as over-sampled and under-sampled datasets [2]. High accuracy was achieved on balanced data; however, performance decreased on the original imbalanced dataset, showing the effect of class imbalance. Another study conducted by Usha and Balaji [3] compared decision tree, naïve Bayes, and k-means algorithms using the same dataset. The results indicated relatively low predictive performance, which was mainly attributed to the absence of comprehensive data preprocessing and feature engineering steps. Several classification methods were tested for employee attrition prediction [4]. Despite the use of cross-validation and train–test splitting, the findings suggested that improved preprocessing could enhance model performance.

METHODOLOGY

A. Dataset

The dataset used here is obtained from the Kaggle website. The objective of this dataset is to identify the factors causing employee attrition, considering demographics, job, and performance. The dataset consists of 1470 instances with 35 attributes, and it does not possess any missing values. The effect of varying attributes of employment on the rate of attrition of the employees is determined using the classification technique in this dataset. IBM data scientists prepared

the dataset used here to discover the reasons for employee attrition, which helps create plans for retaining the employees. This data provides demographics like gender, age, and educational level, and also details of job categories, overtime hours, and earning capacities. These details are explained by attributes mentioned in the following list:

- Attrition – Target Variable, Binary (Yes = 1, No = 0)
- Age – Continuous
- Gender – Nominal (Female, Male)
- Department – Nominal (Sales, Research & Development, Human Resources)
- DistanceFromHome – Discrete Numeric (integer distance units)
- BusinessTravel – Nominal (Non-Travel, Travel_Rarely, Travel_Frequently)
- DailyRate – Continuous
- Education – Ordinal (1 = Below College, 2 = College, 3 = Bachelor, 4 = Master, 5 = Doctor)
- EnvironmentSatisfaction – Ordinal (1 = Low, 2 = Medium, 3 = High, 4 = Very High)
- HourlyRate – Continuous
- JobInvolvement – Ordinal (1 = Low, 2 = Medium, 3 = High, 4 = Very High)
- JobLevel – Ordinal (1 = Entry Level to 5 = Senior Level)
- JobSatisfaction – Ordinal (1 = Low to 4 = Very High)
- MaritalStatus – Nominal (Single, Married, Divorced)
- MonthlyIncome – Continuous
- MonthlyRate – Continuous
- NumCompaniesWorked – Discrete Numeric
- OverTime – Binary (Yes = 1, No = 0)
- PercentSalaryHike – Continuous
- PerformanceRating – Ordinal (3 = Good, 4 = Excellent)
- RelationshipSatisfaction – Ordinal (1 = Low to 4 = Very High)
- StockOptionLevel – Ordinal (0 to 3)
- TotalWorkingYears – Continuous
- TrainingTimesLastYear – Discrete Numeric (number of training sessions)
- WorkLifeBalance – Ordinal (1 = Bad to 4 = Excellent)
- YearsAtCompany – Continuous
- YearsInCurrentRole – Continuous
- YearsSinceLastPromotion – Continuous
- YearsWithCurrManager – Continuous

B. Descriptive Statistics

Descriptive statistics give the first insight into the data by summarizing the data distribution of the numerical variables by looking at the center and dispersion of the data. This is shown by Tables 1 and 2, in which the numerical properties are depicted by Tukey's five-number summary and the mean.

TABLE I. SUMMARY OF NUMERICAL DATA

Variable	min	25%	median	mean	75%	max
Age	18.0	30.0	36.0	36.924	43.0	60.0
DailyRate	102.0	465.0	802.0	802.486	1157.0	1499.0
HourlyRate	30.0	48.0	66.0	65.891	83.75	100.0
MonthlyIncome	1009.0	2911.0	4919.0	6502.931	8379.0	19999.0
MonthlyRate	2094.0	8047.0	14235.5	14313.103	20461.5	26999.0
PercentSalaryHike	11.0	12.0	14.0	15.21	18.0	25.0
TotalWorkingYears	0.0	6.0	10.0	11.28	15.0	40.0
YearsAtCompany	0.0	3.0	5.0	7.008	9.0	40.0
YearsInCurrentRole	0.0	2.0	3.0	4.229	7.0	18.0
YearsSinceLastPromotion	0.0	0.0	1.0	2.188	3.0	15.0
YearsWithCurrManager	0.0	2.0	3.0	4.123	7.0	17.0
DistanceFromHome	1.0	2.0	7.0	9.193	14.0	29.0
NumCompaniesWorked	0.0	1.0	2.0	2.693	4.0	9.0
TrainingTimesLastYear	0.0	2.0	3.0	2.799	3.0	6.0

Table 1 shows descriptive statistics for continuous and discrete numerical variables. Employees' ages range from 18 to 60, with a mean age of 36.92 and a median age of 36; there is a symmetrical age. Monthly income ranges from 1,009 to 19,999, with a mean of 6,502.93 and a median of 4,919; this indicates a right-skewed distribution. Total work experience ranges from 0 to 40 years, with a median of 10 years.

TABLE II. SUMMARY OF ORDINAL NUMERICAL DATA

Variable	min	Q1	median	mean	Q3	max
Education	1.0	2.0	3.0	2.913	4.0	5.0
EnvironmentSatisfaction	1.0	2.0	3.0	2.722	4.0	4.0

JobInvolvement	1.0	2.0	3.0	2.730	3.0	4.0
JobLevel	1.0	1.0	2.0	2.064	3.0	5.0
JobSatisfaction	1.0	2.0	3.0	2.729	4.0	4.0
PerformanceRating	3.0	3.0	3.0	3.154	3.0	4.0
RelationshipSatisfaction	1.0	2.0	3.0	2.712	4.0	4.0
StockOptionLevel	0.0	0.0	1.0	0.794	1.0	3.0
WorkLifeBalance	1.0	2.0	3.0	2.761	3.0	4.0

Table II presents the descriptive statistics of the ordinal variables. The median indicates the position, and the interquartile range denotes the dispersion. The median of 3 for the items of education, job involvement, environment, and work-life balance indicate that the employees view these factors in a moderately to strongly positive light. The IQRs are similar, indicating that the data points are close together and concentrated around these values.

TABLE III. SUMMARY OF CATEGORICAL DATA

Variable	Category	Count
BusinessTravel	Travel Rarely	1043
BusinessTravel	Travel Frequently	277
BusinessTravel	Non-Travel	150
Department	Research & Development	961
Department	Sales	446
Department	Human Resources	63
Gender	Male	882
Gender	Female	588
MaritalStatus	Married	673
MaritalStatus	Single	470
MaritalStatus	Divorced	327
OverTime	No	1054
OverTime	Yes	416

The distribution of different categories within the dataset can be observed by the summary of categorical variables.
There are 588 females and 882 males. Concerning the working conditions, 1,054 of the workers do not perform overtime, and 416 of the workers perform overtime. These categories provide valuable information on the demographics and working arrangements of the workers.

The figure shows a clear positive relationship between age and monthly income for both male and female employees. The regression lines show similar upward trends between genders and indicate that income increases with age regardless of gender. Although male employees appear to earn slightly higher incomes in certain age groups, there is significant overlap between the two groups.

C. Missingness

The dataset used in this study initially contained no missing observations. However, to enable comparison of missing data analysis and imputation methods, synthetic missing values were generated for selected variables.

In the data, approximately 7% of missing values were generated for each of the variables MonthlyIncome, TotalWorkingYears,YearsAtCompany,DistanceFromHome, and JobSatisfaction. Considering the total number of observations in the dataset, the generated missing values correspond to approximately 7% of the entire dataset. The missing values were generated randomly after the data collection process, and the missing observation rate is the same for all relevant variables.

To examine how the missing values are distributed throughout the dataset and to support the exploratory data analysis process, a visualization was performed using a missing value matrix (Figure 1). This visualization shows that the missing values are not concentrated in a particular variable or group of observations, but rather exhibit a scattered and irregular structure throughout the dataset. Furthermore, it was observed that no single observation was missing in all variables, and the missing values did not form a distinct pattern.

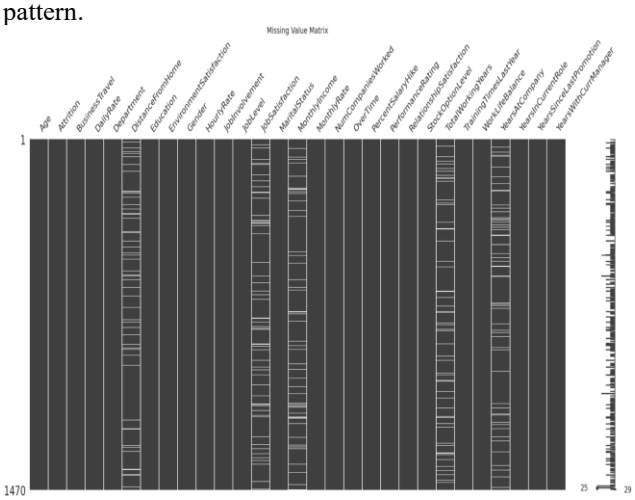

Fig. 1. Missingness Matrix

After the missing values were created, the distributions of the relevant variables were compared before and after the missing values were added. This comparison was performed using both visual and statistical methods. In particular, histograms were examined for the distributions before and after the missing values were added. Visual examinations show that the distribution shapes (e.g., skewness and density regions) are largely preserved.
In addition, the Kolmogorov Smirnov (KS) test was applied to quantitatively compare the distributions. The

test results show that the distributions of the observations obtained after the missing values were added do not differ statistically significantly from the distributions before the missing values were added (p-values > 0.05). This finding supports the idea that the added missing values do not disrupt the fundamental distributional structure of the dataset.

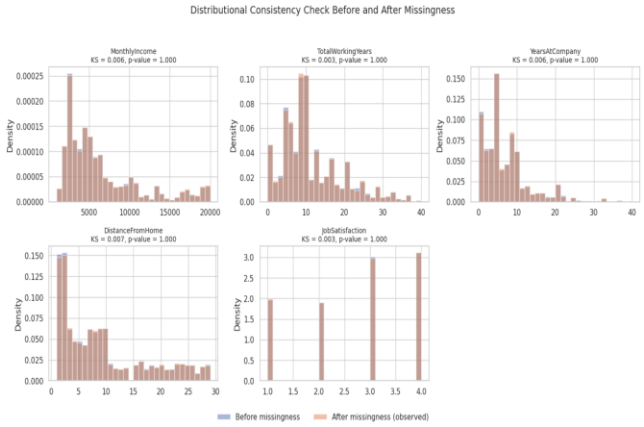

Fig. 2. Distribution Check Plots

TABLE IV.KOLMOGOROV–SMIRNOV STATISTICS FOR SELECTED VARIABLES

Variable	KS_statistic
MonthlyIncome	0.006254
TotalWorkingYears	0.002978
YearsAtCompany	0.006131
DistanceFromHome	0.006525
JobSatisfaction	0.003276

D. Exploratory Data Analysis

In the exploratory data analysis phase, the aim was to understand the basic structure of the dataset, examine the distributions of the variables, and determine the research questions to be tested in the Confirmatory Data Analysis phase. EDA analyses were performed only on the training dataset to prevent potential data leakage.

D.1. Univariate Analysis

D.1.1. Numeric Variable Analysis

Histograms and box plots were examined for the numerical variables Monthly Income, Age, and Total Working Years (Figure 3).

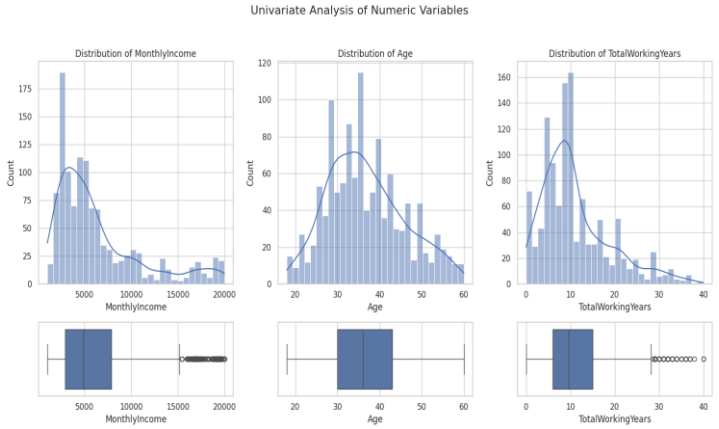

Fig. 3. Univariate Analysis: Histograms and Box Plots of Key Features

The distribution of the Monthly Income variable is significantly right-skewed, with outliers observed at high income levels. The median being lower than the mean indicates that the distribution is not symmetrical. The Age variable exhibits an approximately unimodal distribution, although deviations from a normal distribution were observed. The TotalWorkingYears variable shows a strong right-skewedness, with individuals with particularly high years of service forming outliers.

D.1.2. Categorical Variable Analysis

When examining the distribution of the Job Satisfaction variable, it is observed that a large proportion of employees are concentrated at high satisfaction levels (3 and 4). In contrast, low satisfaction levels (1 and 2) represent a more limited group of employees. This distribution indicates that job satisfaction in the dataset is generally clustered at medium-high levels, and a significant portion of employees are satisfied with their jobs. In the Gender variable, the proportion of male employees (59.5%) is higher than that of female employees (40.5%). These analyses revealed the distribution characteristics and potential imbalances of the variables in the dataset.
(Figure 4).

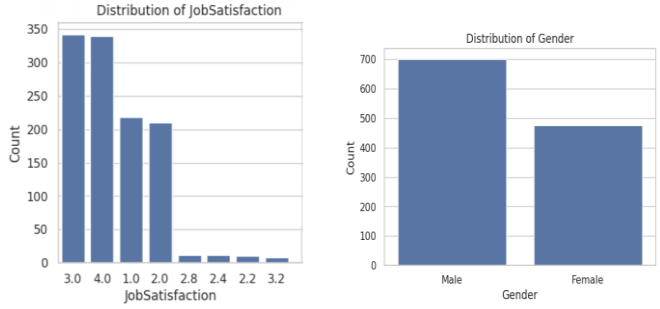

Fig. 4. Distribution of Selected Categorical Variables

D.2. Bivariate Analysis: Target (Attrition) vs Numeric Variables

The relationships between attrition and numerical variables were examined using box plots.
(Figure 5).

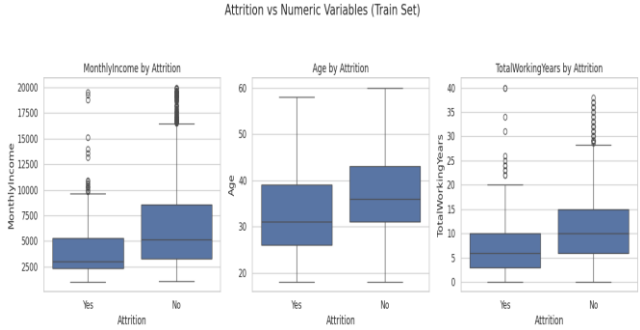

Fig. 5. Box Plots of Numerical Variables by Attrition Status

When the MonthlyIncome variable is examined, it is observed that the average and median monthly income of employees who leave their jobs (Attrition = Yes) is lower compared to employees who remain employed (Attrition = No). In the Age variable, it is observed that employees who left their jobs were concentrated in a younger age group. In terms of the TotalWorkingYears variable, the median total years of employment is lower among employees who leave their jobs.

Based on these visual findings, the following EDA questions were formulated:

RQ1: Are employees with lower monthly incomes more likely to leave their jobs.
RQ2: Are younger employees more likely to leave their jobs compared to older employees?
RQ3: Does the probability of leaving a job decrease as total years of employment increase?

These questions were formulated based solely on observational findings in the EDA part and and were later examined through formal statistical testing in the CDA part.

D.3. Bivariate Analysis: Target (Attrition) vs Categorical Variables

The relationships between attrition and categorical variables were examined using contingency tables, the Chi-square test, and Cramér's V.

TABLE V.CHI-SQUARE TEST RESULTS

JobSatisfaction	Attrition = No	Attrition = Yes
1.0	167	51
1.8	3	0
2.0	176	34
2.2	9	2
2.4	12	0
2.6	4	4
2.8	9	3
3.0	287	55
3.2	7	1
3.4	6	2
3.6	4	0
4.0	302	38

TABLE VI. CHI-SQUARE TEST RESULTS

Statistic	Value
Chi-square statistic	26.350
Degrees of freedom	11
p-value	0.0058
Cramér's V	0.150

A statistically significant relationship was found between Job Satisfaction and Attrition ($p < 0.05$). However, the Cramér's V value (~0.15) indicates a weak to moderate relationship.

TABLE VII. GENDER BY ATTRITION

Gender	Attrition = No	Attrition = Yes
Female	403	73
Male	583	117

TABLE VIII. CHI-SQUARE TEST RESULTS FOR GENDER AND ATTRITION

Statistic	Value
Chi-square statistic	0.302
Degrees of freedom	1
p-value	0.5826
Cramér's V	0.016

No significant relationship was found between the gender variable and attrition ($p > 0.05$, Cramér's $V \approx 0.02$), indicating that gender does not play a strong role in explaining job turnover in this dataset.

TABLE IX. OVERTIME BY ATTRITION

OverTime	Attrition = No	Attrition = Yes
No	757	84
Yes	229	106

Table VIII. OverTime by Attrition

TABLE X. CHI-SQUARE TEST RESULTS

Statistic	Value
Chi-square statistic	81.333
Degrees of freedom	1
p-value	0.0000
Cramér's V	0.263

A statistically significant and relatively stronger relationship was observed between the OverTime variable and attrition ($p < 0.001$, Cramér's $V \approx 0.26$).

D.4. Multivariate Structure

The relationships between selected numerical variables (Monthly Income, Age, Total Working
Years, Years At Company, Distance From Home) were examined using a correlation heatmap (Figure 6).

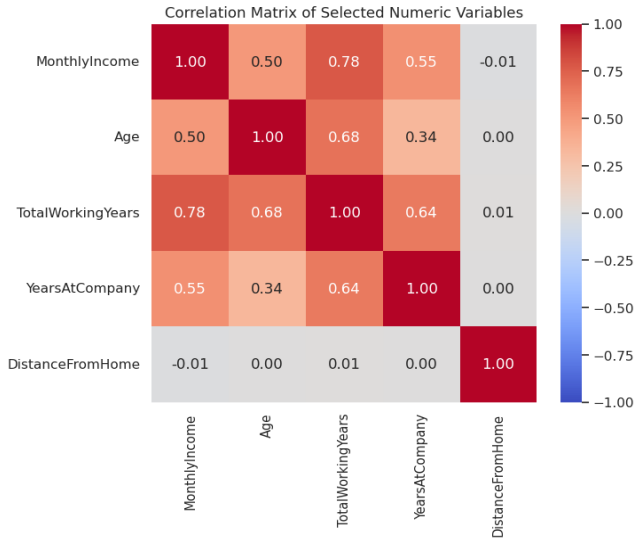

Fig. 6. Correlation Matrix Heatmap of Selected Numerical Variables

Moderately strong level of positive correlation ($r \approx 0.78$) was observed between Monthly Income and Total Working Years.There is also moderately strong positive relationship between Age and Total Working Years ($r \approx 0.68$). A weak positive correlation ($r \approx 0.34$) was observed between Age and Years At Company. The Distance From Home variable was found to have no significant correlation with the other variables.This structure formed the basis for dimensionality reduction and modeling steps in the later stages.

E. Confirmatory Data Analysis

In this part of study firstly, assumption checks are conducted as normality and homogeneity of variance assumptions.Since the Attrition variable has two categories (Yes / No), the distribution characteristics of the numerical variables were examined separately for each group. Tests were performed for the Monthly Income, Age, and Total Working Years variables.

E.1. Normality Assumption

The normality assumption was evaluated for each numerical variable in the Attrition = Yes and Attrition = No groups using the Shapiro–Wilk test and Q–Q (Quantile–Quantile) plots.

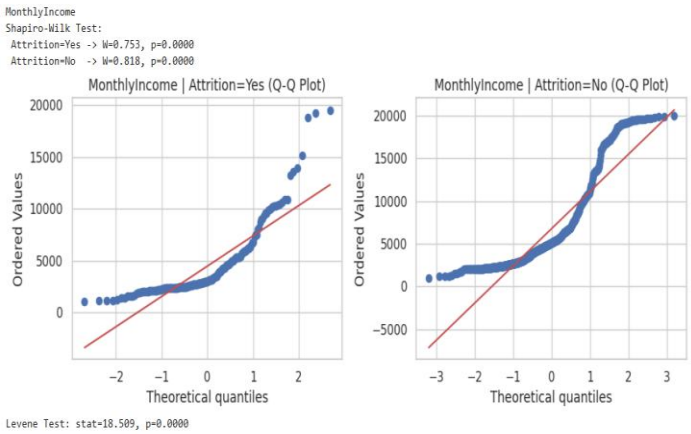

Fig. 7. Normality Assumption Check for Monthly Income (Q-Q Plots)

For the MonthlyIncome variable, the Shapiro–Wilk test yielded significant results in both groups ($p < 0.05$). In the Q–Q plots, significant deviations were observed, particularly in the upper tails, indicating a right-skewed structure for income distribution. These findings suggest that the MonthlyIncome variable does not satisfy the normal distribution assumption for either group.

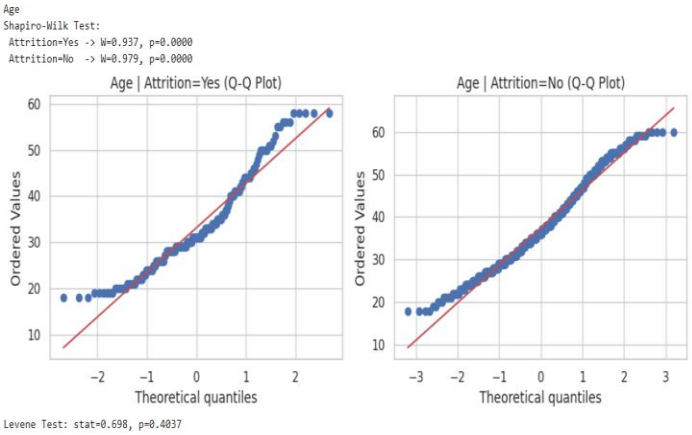

Fig. 8. Normality Assumption Check for Age (Q-Q Plots)

For the Age variable, the Shapiro–Wilk test was significant for both groups ($p < 0.05$). In the Q–Q plots, the distribution was observed to be close to normal in the middle region, but deviations from linearity were seen, particularly at the extremes. This indicates that the age variable does not fully conform to a normal distribution.

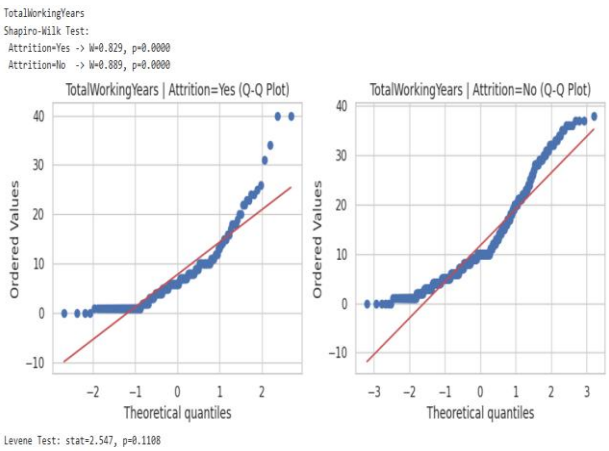

Fig. 9. Normality Assumption Check for Total Working Years (Q-Q Plots)

The Shapiro-Wilk test results are significant for both groups in the TotalWorkingYears variable ($p < 0.05$). Significant deviations are present in the Q-Q plots, particularly at low and high values. This finding indicates that the total working years variable also does not meet the assumption of normal distribution.

Overall, it was concluded that none of the three numerical variables examined met the assumption of normal distribution based on the Attrition groups.

E.2. Assumption of Homogeneity of Variances

The equality of variances between groups was evaluated using the Levene test.

The Levene test for the MonthlyIncome variable is statistically significant ($p < 0.05$). This result indicates that the variances of the Attrition = Yes and Attrition = No groups are not equal, and the assumption of homogeneity of variances is violated (Figure 7).

The Levene test for the Age variable is not significant ($p > 0.05$). This finding suggests that the variances between groups can be considered homogeneous in terms of the age variable(Figure 8).

The Levene test is also not significant for the TotalWorkingYears variable ($p > 0.05$), and the assumption of homogeneity of variance is met(Figure 9).

Since the normality assumption is violated for all numerical variables and homogeneity of variance is not met for the MonthlyIncome variable, the use of parametric tests is not appropriate. Therefore, the non-parametric Mann–Whitney U test was preferred to examine the differences between the Attrition = Yes and Attrition = No groups.

E.3. Mann–Whitney U Test Results

As a result of the assumption checks performed in the previous step, it was observed that the variables MonthlyIncome, Age, and TotalWorkingYears did not meet the normality assumption for both Attrition groups (Shapiro–Wilk tests, all p-values < 0.05). Furthermore, the homogeneity of variances assumption was not met for some variables. Therefore, instead of parametric tests, the Mann–Whitney U test, a non-parametric method, was preferred for intergroup comparisons.

TABLE XI. MANN–WHITNEY U TEST RESULTS

Variable	Median (Attrition = Yes)	Median (Attrition = No)	U statistic	p-value
MonthlyIncome	3007.0	5128.0	60116.0	4.975272e-15
Age	31.0	36.0	66119.5	1.268360e-10
TotalWorkingYears	6.0	10.0	61382.5	4.461901e-14

The Mann–Whitney U test result for the Monthly Income variable is statistically significant ($p < 0.001$). The median monthly income of the Attrition = Yes group (3007) is significantly lower than that of the Attrition = No group (5128). This finding statistically supports the hypothesis formulated in the EDA part that employees with lower monthly incomes are more likely to leave their jobs.

The Mann–Whitney U test result for the Age variable is also statistically significant ($p < 0.001$). The median age of employees who left their jobs (31) is lower than the median age of employees who remained employed (36). This result confirms the EDA finding that younger employees are more likely to leave their jobs.

The Mann–Whitney U test also revealed a statistically significant difference in terms of the Total Working Years variable ($p < 0.001$). The median total working years in the Attrition = Yes group is 6, while in the Attrition = No group it is 10. This result is consistent with the hypothesis that the likelihood of leaving a job decreases as the total number of years worked increases.

The results of the Mann–Whitney U test show that all three main questions observationally found in the EDA part are statistically supported. Thus, it has been confirmed by confirmatory analyses that the variables of monthly income, age, and total years worked are significantly related to employees' attrition behavior.

E.4. Tests and Results for Questions in EDA

Based on the observational findings obtained in the EDA part, the following directional hypotheses were tested:

H1: The monthly income of employees who left the company (Attrition = Yes) is lower than that of employees who remained in the company (Attrition = No).
H2: Employees who left the company are younger than those who remained in the company.
H3: The total years of service of employees who left the company are lower than those who remained in the company.

Since the normality assumption was not met according to the Shapiro–Wilk test and Q-Q plots, the Mann–Whitney U test was preferred instead of parametric tests. Because the hypotheses established in the EDA part were one-sided, the tests were applied as one-sided.

TABLE XII. MANN–WHITNEY U TEST RESULTS

Variable	Median (Yes)	Median (No)	U statistic	p-value (one-sided, Yes<No)	p-value (two-sided)	Effect size (Rank-biserial r)	n(Yes)	n(No)
MonthlyIncome	3007.0	5128.0	60116.0	2.487636e-15	4.975272e-15	0.358215	190	986
TotalWorkingYears	6.0	10.0	61382.5	2.230950e-14	4.461901e-14	0.344694	190	986
Age	31.0	36.0	66119.5	6.341800e-11	1.268360e-10	0.294123	190	986

According to the Mann–Whitney U test results, the monthly income of employees who left the company (Attrition = Yes) is statistically significantly lower than that of employees who remained in the company (Attrition = No) ($p < 0.001$).

Furthermore, the calculated effect size (rank-biserial $r \approx 0.36$) indicates that this difference has a moderate effect ($p < 0.001$). It was found that the monthly income of employees who left the company is significantly lower than that of employees who remained in the company.

The median age of employees who left the company (31) is significantly lower than the median age of employees who remained in the company (36) ($p < 0.001$). The calculated effect size value ($r \approx 0.29$) indicates that this difference has a weak to moderate effect.

It was concluded that employees who left the company are significantly younger than those who remained in the company.

In terms of total years of service, it was observed that employees who left their jobs had significantly fewer total years of service than those who remained employed ($p < 0.001$).
The calculated effect size ($r \approx 0.34$) indicates that this difference corresponds to a moderate effect size.

F. Feature Engineering & Dimension Reduction

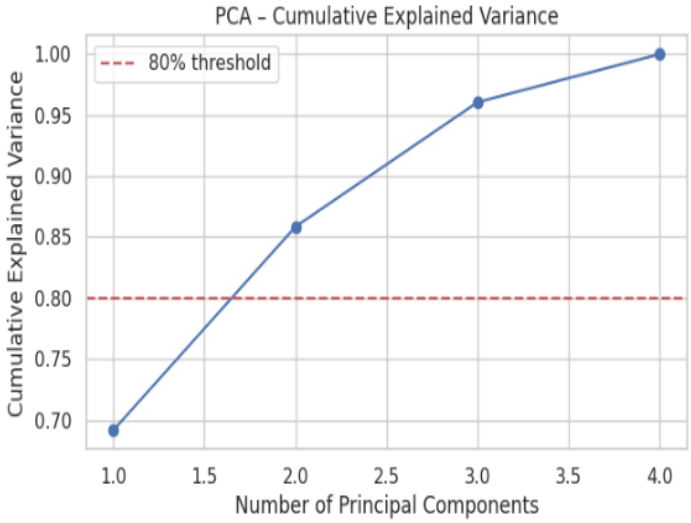

Fig. 10. PCA Cumulative Explained Variance Plot

TABLE XIII. PRINCIPAL COMPONENT LOADINGS FOR PC1 AND PC2

Variable	PC1	PC2
MonthlyIncome	0.518197	-0.093435
Age	0.451415	0.726761
TotalWorkingYears	0.567459	0.049612
YearsAtCompany	0.453526	-0.678695

PCA Train Shape: (1176, 2)
PCA Test Shape: (294, 2)

Nearly 86% of the variability in the chosen numerical features can be described by the first two principal components, according to the cumulative explained variance analysis. This indicates that significant dimensionality reduction is possible without significant information loss. A general idea of employee experience and compensation is reflected in the first principal

component (PC1), which is strongly correlated with MonthlyIncome, Age, TotalWorkingYears, and YearsAtCompany. The second principal component (PC2) mainly reflects the relationship between employees' age and how long they have been with the company. Age and time spent in the company indicate different directions; being older does not necessarily mean having worked longer in the organization. These findings are consistent with the patterns observed during the exploratory data analysis and suggest that various experience-related variables carry similar information. Representing the data with a smaller set of uncorrelated components is a more concise and efficient approach than using the original variables for later modeling.

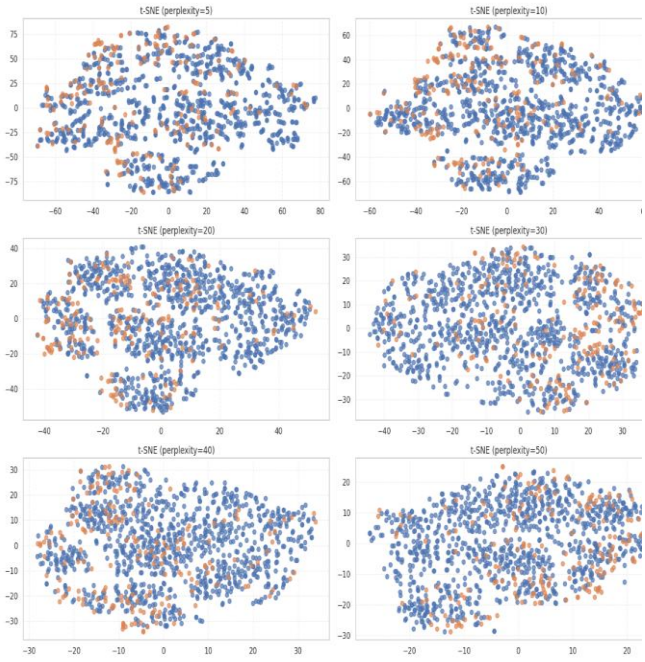

Fig. 11. t-SNE Visualization with Various Perplexity Values

t-SNE was used as an exploratory visualization technique on the standardized training data to explore whether employees with different attrition outcomes show any natural separation in a low-dimensional space. Nevertheless, in all these contexts, it has been found that there is a significant level of overlap between the two attrition groups. This overlap means that when the data are shown in fewer dimensions, the two groups cannot be clearly separated from each other.

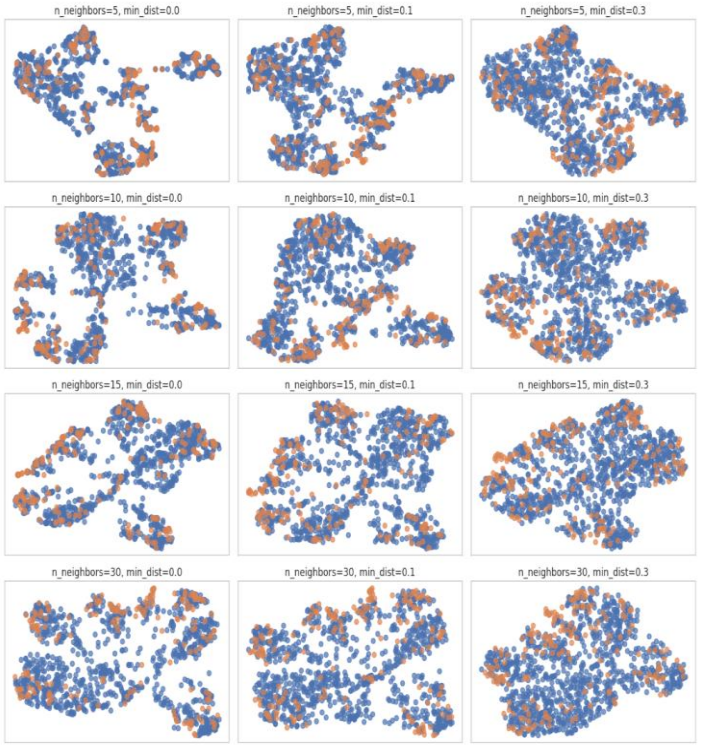

Fig. 12. UMAP Visualization with Various Hyperparameters

UMAP was employed to explore the data to a greater extent with varying neighborhood numbers and a range of minimum distances. Although patterns were fairly robust across varying parameters, a significant amount of overlap amongst attrition classes was noticed. This reveals that the issue of employee attrition is affected by diverse and interlinked factors, as there are no logically separate categories associated with it.

G. Statistical Modeling (Prediction/Classification)

G.1. Experimental Setup & Preprocessing

The dataset was split into training and test sets using stratified sampling. This preserved the class distribution of the binary target variable (Attrition). A fixed random seed was used to ensure reproducibility. Model selection and comparison were performed using five-fold stratified cross-validation. This step was applied only to the training set. The test set was kept separate for the final and unbiased evaluation.

All preprocessing steps were applied using pipelines to prevent data leakage. Numerical variables were imputed using the median and then standardized. Categorical variables were imputed using the most frequent category. They were encoded using one-hot encoding. All preprocessing steps were fitted only on the training

data. The same transformations were then applied to the validation and test sets.

For models that used oversampling, SMOTE was included inside the cross-validation pipeline. Synthetic samples were generated only from the training folds. This prevented artificial information from entering the validation or test data.

G.2. Logistic Regression (Base Model)

Logistic regression was chosen as the baseline model since the response variable is binary. The mean cross-validated ROC-AUC for the logistic regression was approximately 0.83. The model properly identified some attrition cases, achieving a recall of about 0.40 for the minority class on the test set. The model had a high specificity of about 0.96. This indicates strong performance in identifying non-attrition cases. A second logistic regression model was trained using PCA features. Components that explained 80% of the data were retained by the PCA. The performance of this model was poorer. The cross-validated ROC-AUC decreased to roughly 0.68. The model found no instances of attrition in the test set. For the minority class, the recall was zero. Although the specificity was 1.00, this outcome is deceptive. Because the data is unbalanced, the model typically predicted the majority class. As a result, the PCA-based model was ineffective for predicting attrition. It was only retained for comparison.

G.3. Regularization (Lasso & Ridge)

To reduce multicollinearity and model complexity, regularized logistic regression models (Ridge, Lasso, and Elastic Net) were tested. In these models five-fold stratified cross-validation was applied. As the evaluation metric ROC-AUC was used . The performance of the models was found to be similar to the basic logistic regression model. ROC-AUC values are approximately 0.83. Sensitivity values for minority class range from approximately 0.40 to 0.43. Only minor differences were observed between the models. Regularization slightly altered the balance between recall and specificity. However, the recall for the minority class and the F1-score did not significantly improve. The issue of class imbalance remained. As a result, regularization alone was not enough for this dataset.

G.4. Handling Class Imbalance (SMOTE & Cost-Sensitive Learning)

There is a significant imbalance in the target variable. Therefore, specific methods have been applied to overcome this imbalance. Two approaches are applied in this section. These are SMOTE with oversampling and class-weighted logistic regression with cost-sensitive learning.

The SMOTE-based logistic regression model showed an increase in Sensitivity (approximately 0.74) compared to other models. But there is a decrease in Specificity (approximately 0.77). SMOTE captured the minority class (attrition = Yes) better. On the other hand, the overall stability of the model and its success in correctly distinguishing the negative class (attrition = No) decreased.
The cost-sensitive logistic regression model gives the highest Sensitivity (approximately 0.79) among all models. It also maintained a reasonable specificity (approximately 0.77). The cross-validated ROC-AUC value was slightly lower than the base and modified models. But the improvement in minority class detection was a more significant change. The cost-sensitive model achieved the highest F1 score among the compared approaches. This indicates a more balanced compromise between sensitivity and recall.

H. Model Evaluation & Conclusion

TABLE XIV . PERFORMANCE COMPARISON OF CLASSIFICATION MODELS

Model	CV_ROC_AUC_Mean	CV_ROC_AUC_Std	Recall (Sensitivity)	Specificity	F1	ROC_AUC	Kappa
Logistic (Imbalanced baseline)	0.829856	0.038065	0.404255	0.963563	0.506667	0.830821	0.439798
Ridge (L2)	0.829374	0.037795	0.425532	0.955466	0.512821	0.832458	0.441902
Elastic Net	0.828920	0.037676	0.404255	0.955466	0.493506	0.832113	0.421435
Lasso (L1)	0.827532	0.038582	0.425532	0.959514	0.519481	0.832371	0.451105
Cost-sensitive (class_weight=balanced)	0.817532	0.036100	0.787234	0.769231	0.524823	0.834525	0.396100
SMOTE + Logistic	0.814861	0.032238	0.744681	0.773279	0.507246	0.830649	0.375601
PCA (80%) + Logistic	0.681804	0.035098	0.000000	1.000000	0.000000	0.628047	0.000000

This study compares various statistical classification models suitable for unbalanced data structures. Sensitivity, Specificity, F1-Score, ROC-AUC, and Cohen's Kappa metrics were used in the evaluation.

Basic logistic regression models and regularized models (Ridge, Lasso, and Elastic Net) yielded relatively high ROC-AUC values. However, the sensitivity values of these models were quite low. This indicates that a large portion of employees who left their jobs were not accurately identified. Although the specificity values were high, these models mostly predicted the majority class.

The PCA based logistic regression model performed quite poorly. The model predicted all observations as majority (No) class. Therefore, while the sensitivity and F1-score were zero, the specificity value was equal to one. This result shows that PCA eliminates the necessary information to separate the status of employees who left their jobs.

Balanced modeling approaches yielded more successful results in identifying the minority class. The model using SMOTE significantly increased sensitivity. However, this increase was accompanied by lower specificity and lower kappa values. This indicates a decrease in overall classification fit.

The cost-sensitive logistic regression model achieved the highest sensitivity and F1 score among all models. The ROC–AUC value remained at a competitive level. The balance between sensitivity and specificity is acceptable. Furthermore, this model more effectively identified employees at risk of resignation without generating synthetic data.

In conclusion, the cost-sensitive logistic regression model offered the most balanced and meaningful performance for this unbalanced classification problem and was selected as the final model.

TABLE XV. LOGISTIC REGRESSION COEFFICIENTS AND ODDS RATIOS

Intercept: −0.1775726		
Variable	**Beta**	**Odds_Ratio**
BusinessTravel_Travel_Frequently	0.729067	2.073146
YearsSinceLastPromotion	0.568335	1.765325
OverTime (No)	0.518594	1.679664
NumCompaniesWorked	0.480376	1.616682
DistanceFromHome	0.353779	1.424441
MaritalStatus (Single)	0.311771	1.365841
YearsAtCompany	0.203937	1.226221
PerformanceRating	0.121499	1.129188
Department_Sales	0.083175	1.086732
MonthlyRate	0.018450	1.018621
JobLevel	0.017288	1.017438
Education	0.001238	1.001239
HourlyRate	-0.005462	0.994553
Department_Human Resources	-0.014341	0.985762
PercentSalaryHike	-0.149774	0.860903
DailyRate	-0.160102	0.852057
BusinessTravel_Travel_Rarely	-0.168037	0.845322
Gender (Male)	-0.183886	0.832031
TrainingTimesLastYear	-0.194595	0.823168
TotalWorkingYears	-0.211387	0.809461
StockOptionLevel	-0.223563	0.799665
WorkLifeBalance	-0.265835	0.766655
RelationshipSatisfaction	-0.294201	0.745127
YearsInCurrentRole	-0.359036	0.698349
JobInvolvement	-0.359845	0.697785
JobSatisfaction	-0.403870	0.667731
MaritalStatus (Married)	-0.457549	0.632833
EnvironmentSatisfaction	-0.460524	0.630953
Age	-0.492454	0.611125
YearsWithCurrManager	-0.516733	0.596466
Gender (Female)	-0.528555	0.589456
MaritalStatus (Divorced)	-0.566663	0.567416
MonthlyIncome	-0.568871	0.566164
Intercept	-0.712441	0.490445
Department_Research & Development	-0.781277	0.457821
OverTime (No)	-1.231035	0.291990
BusinessTravel_Non-Travel	-1.273471	0.279859

Generally there are reasonable coefficients for cost-sensitive logistic regression models. The signs of the coefficients make sense. Positive coefficients increase the chance of leaving the job, while negative coefficients decrease.

Frequent business travel, overtime, distance from work, and a long period without promotion have positive coefficient. So this increase the risk of leaving the job. These variables are related to workload and career dissatisfaction.

On the other hand, as monthly income increases, the probability of leaving the job decreases. Job satisfaction, environmental satisfaction also have negative coefficient. Increasing the length of time spent with the current manager and in the company also reduces the risk of leaving. These results are the expected relationships in terms of employee commitment.
Multicollinearity for numerical variables has been controlled using regularization methods. And the model assumptions are mostly met. The log-odds linear assumption is generally reasonable.

REFERENCES

[1] Setiawan, I., Suprihanto, S., Nugraha, A. C., & Hutahaean, J. (2020). HR analytics: Employee attrition analysis using logistic regression. IOP Conference Series: Materials Science and Engineering, 830, 032001. https://doi.org/10.1088/1757-899X/830/3/032001

[2] S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 2018, pp. 93-98, doi: 10.1109/INNOVATIONS.2018.8605976.

[3] Usha, P. M., and Balaji, N. V. (2021). A comparative study of machine learning algorithms for employee attrition prediction. Iop Conference Series Materials Science and Engineering. doi: 10.1088/1757-899x/1085/1/012029

[4] Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020, November 3). Predicting employee attrition using machine learning techniques. MDPI. https://www.mdpi.com/2073-431X/9/4/86