IBM HR Analytics Employee Attrition & Performance

Name: Raj Purohith Arjun UIN:535005964

1.Overview of the Project

The IBM HR Analytics Employee Attrition & Performance dataset provides insights into factors influencing employee attrition. By analysing employee demographics, job satisfaction, work-life balance, and performance metrics, this project aims to identify the key contributors to attrition, helping organizations implement data-driven strategies to enhance retention, job satisfaction, and overall performance.

Objective:

The primary objective of this project is to analyze the factors influencing employee attrition within an organization by examining variables such as employee age, gender, monthly income, distance from home, job role, and various satisfaction levels. Specifically, we aim to investigate how factors like job satisfaction, environment satisfaction, work-life balance, and performance ratings contribute to the likelihood of employees leaving the organization. Using statistical and machine learning techniques, we will identify significant predictors of attrition and develop a model to estimate the probability of an employee attriting. This analysis will provide actionable insights for improving employee retention and optimizing organizational performance.

Research Questions:

What demographic and workplace characteristics are most strongly associated with employee attrition?

How do income and distance from home influence attrition probability?

Can statistical or machine learning models accurately predict attrition?

Are some job roles or departments more prone to attrition?

How do performance ratings affect retention across different satisfaction levels?

2. Dataset Description

2.1 Source of the Dataset

The dataset was sourced from the IBM HR Analytics Employee Attrition & Performance repository available on Kaggle.

Link: IBM HR Analytics Employee Attrition & Performance

2.2 Description of Variables

This dataset focuses on employee information and workplace attributes, with the goal of analyzing employee attrition, a key factor in understanding why employees leave the organization. By studying the various factors affecting attrition, the goal is to develop strategies to improve employee retention and organizational performance.

Dataset Characteristics:

- **Rows (Observations)**: 1,470 employees
- **Columns (Variables)**: The dataset contains 35 variables. Below is a full list of the variables included:

Variables Description

Dependent Variable (Target):

- **Attrition**: This is the target variable, indicating whether the employee has left the organization (Yes = Attrited, No = Not Attrited).

Independent Variables:

These variables are used to predict or analyze the likelihood of attrition:

1. **Age**: The age of the employee (numeric).
2. **Gender**: The employee's gender (Male/Female).
3. **BusinessTravel**: Frequency of business travel (categorical).
4. **DailyRate**: Daily rate of the employee (numeric).
5. **DistanceFromHome**: Distance between the employee's home and workplace (numeric).
6. **Education**: The highest level of education completed (numeric: 1 to 5).
7. **EducationField**: Field of education (Life Sciences, Medical, Other, Marketing, Technical Degree).
8. **EmployeeCount**: Number of employees in the company (constant).
9. **EmployeeNumber**: Unique identifier for each employee (constant).
10. **EnvironmentSatisfaction**: Satisfaction with the work environment (ordinal).
11. **HourlyRate**: Hourly rate of the employee (numeric).
12. **JobInvolvement**: Employee's involvement in their job (ordinal).
13. **JobLevel**: Job level (numeric).
14. **JobRole**: The role or job title of the employee (Sales Executive, Research Scientist, etc.).
15. **JobSatisfaction**: Employee's satisfaction with their job (ordinal).
16. **MaritalStatus**: Marital status of the employee (Single, Married, Divorced).
17. **MonthlyIncome**: Monthly income of the employee (numeric).
18. **MonthlyRate**: Monthly rate of the employee (numeric).
19. **NumCompaniesWorked**: Number of companies the employee has worked for (numeric).
20. **OverTime**: Whether the employee works overtime (categorical: Yes, No).
21. **PercentSalaryHike**: Percentage salary increase (numeric).
22. **PerformanceRating**: Performance rating of the employee (ordinal).

23. **RelationshipSatisfaction**: Satisfaction with workplace relationships (ordinal).
24. **StandardHours**: Standard hours (constant).
25. **StockOptionLevel**: Level of stock options (numeric).
26. **TotalWorkingYears**: Total years of work experience (numeric).
27. **TrainingTimesLastYear**: Number of training sessions in the past year (numeric).
28. **WorkLifeBalance**: Employee's work-life balance (ordinal).
29. **YearsAtCompany**: Number of years the employee has worked at the company (numeric).
30. **YearsInCurrentRole**: Number of years in the current role (numeric).
31. **YearsSinceLastPromotion**: Number of years since last promotion (numeric).
32. **YearsWithCurrManager**: Number of years with the current manager (numeric).

Irrelevant Variables:

1. **EmployeeCount**: Constant value across rows, irrelevant to attrition analysis.
2. **StandardHours**: Constant value for all employees, adds no value to prediction.
3. **EmployeeNumber**: Unique identifier, irrelevant to attrition behavior. prediction.

3.2 Encoding Categorical Variables

3.2.1 Label Encoding for Ordinal Variables:

Ordinal variables like EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance, PerformanceRating , RelationshipSatissfaction and WorkLifeBalance can be encoded using Label Encoding.

Ordinal Variables:

EnvironmentSatisfaction: Ordered levels (e.g., "Low", "Medium", "High", "Very High").

JobInvolvement: Ordered levels (e.g., "Low", "Medium", "High").

JobSatisfaction: Ordered levels (e.g., "Low", "Medium", "High", "Very High").

PerformanceRating: Ordered levels (e.g., "Excellent", "Good", "Needs Improvement").

RelationshipSatisfaction: Ordered levels (e.g., "Low", "Medium", "High").

WorkLifeBalance: Ordered levels (e.g., "Bad", "Good", "Best").

- Purpose: The goal of label encoding is to convert ordinal categorical variables (those with a meaningful order or ranking) into integer values. This transformation ensures that the relative ordering of categories is preserved.
- Columns: The columns selected for label encoding include ['EnvironmentSatisfaction', 'JobInvolvement', 'JobSatisfaction', 'PerformanceRating', 'RelationshipSatisfaction', 'WorkLifeBalance'], all of which represent factors that can be ordered or ranked (e.g., Low, Medium, High).

- Method: We used LabelEncoder to transform these categorical values into integers, where categories like Low, Medium, and High were mapped to numerical values (e.g., Low = 0, Medium = 1, High = 2). This ensures the ordinal nature of the data is represented numerically.
- Result: After applying label encoding, the ordinal columns are now represented as numeric values, allowing them to be used effectively in machine learning models that require numeric input.

3.2.2 One-Hot Encoding for Nominal Variables:

Nominal Variables:

BusinessTravel: Categories (e.g., "Travel_Frequently", "Non-Travel") have no inherent order.

Department: Different departments (e.g., "Sales", "HR") with no rank or order.

EducationField: Different fields (e.g., "Life Sciences", "Marketing") without order.

Gender: Categories ("Male", "Female") with no ranking.

JobRole: Different job titles (e.g., "Manager", "Sales Executive") with no inherent order.

MaritalStatus: Categories ("Single", "Married") with no order.

OverTime: Binary categories ("Yes", "No") with no ranking.

StockOptionLevel: Different levels (e.g., "0", "1") representing categories, not a numeric scale.

Purpose: One-hot encoding is used to convert nominal categorical variables (those without any specific order) into binary columns. This technique is useful because many machine learning algorithms require numeric input, and one-hot encoding avoids assigning an arbitrary order to categorical values.

Columns: The columns selected for one-hot encoding include ['BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'OverTime', 'StockOptionLevel'], where each represents a category without any inherent ranking (e.g., Gender: Male, Female).

Method: We used pd.get_dummies() to create binary columns for each category in the nominal variables. For instance, for the 'BusinessTravel' column, new columns such as BusinessTravel_Travel_Frequently were created. Each binary column indicates whether an employee frequently travels (1 if true, 0 if false).

Result: The nominal columns are now expanded into multiple binary columns, each representing a specific category, making it suitable for input into machine learning models.

3.3 Scaling and Normalization

Standardization (Z-Score):

The StandardScaler was applied to the numerical columns ['MonthlyIncome', 'MonthlyRate']. This technique transforms the data so that each feature has a mean of 0 and a standard deviation of 1, making the data comparable across different scales.

Transformation: The scaler fits the data and then transforms the values. After transformation, each value in these columns represents how many standard deviations it is away from the mean of that feature.

Output:

The first few rows show the scaled values for MonthlyIncome and MonthlyRate. For example, a value of -0.108350 for MonthlyIncome means that the original value is slightly below the mean of that feature. Similarly, 0.726020 for MonthlyRate indicates that it is above the mean, but less than one standard deviation.

Shape of the Dataset: The shape of the dataset is (1470, 35), meaning that after transformation, the dataset still contains 1470 rows and 35 columns, with only the numerical columns being scaled.

4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

4.1.1 Summary Statistics

Summary statistics provide a numerical snapshot of the dataset's characteristics, aiding in understanding central tendencies, variability, and distribution. Key metrics include:

Mean: Average value of a variable, indicating central tendency.

Median: Midpoint value, robust against outliers and skewed data.

Mode: Most frequently occurring value, revealing common traits.

Standard Deviation: Measure of data spread around the mean.

Minimum and Maximum: Indicate data boundaries and potential outliers.

Percentiles (Q1, Q3): Divide data into lower and upper quartiles, identifying distribution spread.

Key observations include:

Insights from Summary Statistics:

- **Age**:
    - Mean age is approximately 37 years, with a wide range from 18 to 60 years.
    - Indicates a diverse workforce in terms of age distribution.
- **DailyRate**:
    - Average daily rate is 802, with substantial variability (std ~ 403).
    - Suggests a significant difference in daily pay among employees.
- **DistanceFromHome**:
    - Employees live an average of 9.19 units (e.g., miles) from work, with a range of 1 to 29 units.
    - This could influence work-life balance or commuting patterns.
- **Education**:

    - Median education level is 3, suggesting most employees have a bachelor's or similar level of education.
    - Spans from 1 (lowest) to 5 (highest), indicating a mix of educational backgrounds.
- **EnvironmentSatisfaction**:
    - Average satisfaction is 1.72 (scale likely 0-3).
    - Shows room for improvement in workplace satisfaction.
- **HourlyRate**:
    - Mean hourly rate is approximately 66, ranging from 30 to 100.
    - Highlights variability in compensation.
- **JobInvolvement**:
    - Average score is 1.73 (likely scale 0-3), indicating moderate involvement.
    - Potential area for initiatives to increase engagement.
- **JobLevel**:
    - Median level is 2, with a range from 1 to 5.
    - Reflects a typical hierarchical structure.
- **JobSatisfaction**:
    - Average satisfaction is 1.73 (scale likely 0-3).
    - Similar to environment satisfaction, opportunities for improvement exist.
- **MonthlyIncome and MonthlyRate**:
    - Both variables have been standardized (mean ~ 0, std ~ 1).
    - Direct interpretation of raw values is not meaningful without context.
- **NumCompaniesWorked**:
    - Employees have worked at an average of 2.7 companies, with a wide range (0-9).
    - Reflects varying levels of industry experience.
- **PercentSalaryHike**:
    - Average increase is 15%, with a range of 11% to 25%.
    - Reflects annual appraisal increments.
- **RelationshipSatisfaction**:
    - Similar trends to job satisfaction; room for improvement exists.
- **StockOptionLevel**:
    - Median level is 1, with some employees receiving higher stock options.
    - Indicates variation in benefits offered.
- **TotalWorkingYears**:

    - Mean total years worked is 11.28, with a range of 0 to 40.
    - Suggests a mix of experienced and early-career employees.
- **TrainingTimesLastYear**:
    - Median is 3, indicating moderate training frequency.
    - Variability shows differences in training opportunities.
- **WorkLifeBalance**:
    - Average score is 1.76 (scale likely 0-3).
    - Highlights a potential area for improvement in employee well-being.
- **YearsAtCompany**:
    - Median tenure is 5 years, ranging from 0 to 40.
    - Reflects varying degrees of employee loyalty and tenure.
- **YearsInCurrentRole**:
    - Median is 3 years, with some employees holding roles for up to 18 years.
    - Indicates a mix of recently and long-tenured role holders.
- **YearsSinceLastPromotion**:
    - Median is 1 year, but some employees haven't been promoted for up to 15 years.
    - Reflects a potential issue with career progression for some employees.
- **YearsWithCurrManager**:
    - Median is 3 years, indicating moderate stability in reporting relationships.

positions.

4.1.2 Visualization

Histograms, Box Plots, and Density Plots are essential to visualize distributions, detect outliers, and identify patterns in key variables.

1.Histograms

Key Insights and Findings

Demographic Profile
The age distribution shows a bell-shaped curve centered around 30-35 years, indicating a predominantly young to middle-aged workforce. Most employees have an education level of 3.0, suggesting a well-educated workforce, with another significant group at level 4.0.

Work Environment Characteristics

Distance and Location
The distance-from-home histogram reveals that a majority of employees live within 10 units of their workplace, with the highest concentration within 0-5 units. This suggests a locally concentrated workforce.

Work Experience and Tenure
Total working years peak around 10-15 years, indicating a mid-career dominant workforce.
Years at company shows a concerning trend with most employees having less than 5 years of tenure, potentially signaling high turnover.

Professional Development
The training frequency histogram shows that most employees received 2-3 training sessions in the last year. This moderate level of professional development investment could be enhanced to improve retention.

Satisfaction Metrics

Job Satisfaction
The distribution shows an interesting bimodal pattern with peaks at levels 1 and 3, suggesting a polarized workforce in terms of job satisfaction. This could be a critical indicator for potential attrition risks.

Work-Life Balance
The work-life balance metric shows a strong concentration around levels 2 and 3, indicating generally positive perceptions among employees.

Compensation Structure
The monthly income histogram displays a right-skewed distribution, with a large concentration of employees at the lower end of the pay scale. This salary distribution pattern, combined with the satisfaction metrics, could be a contributing factor to potential attrition.

These insights suggest several areas requiring attention, particularly around employee retention strategies, compensation structure, and job satisfaction improvement initiatives. ng levels of satisfaction across different job aspects. g levels of satisfaction across different job aspects.

2.Box Plots

Key Insights and Findings

Distribution Analysis

Age Distribution
The boxplot shows a median age around 35 years, with the interquartile range spanning roughly from 30 to 40 years. The whiskers extend from approximately 20 to 60 years, indicating a well-distributed age range with no significant outliers.

Distance From Home
A right-skewed distribution is evident with the median around 7-8 units. The box (IQR) shows most employees live within 5-15 units from work, while the extended upper whisker indicates some employees commute from up to 30 units away.

Education and Job Satisfaction
Education levels show a compact box between levels 2-4, suggesting consistent educational qualifications across the workforce.
Job satisfaction displays an even distribution across all four levels (0-3), with the box spanning almost the entire range.

Career Metrics

Monthly Income
The income boxplot reveals significant right skewing with numerous high-income outliers. The main body of the distribution (box) is concentrated in the lower income ranges, suggesting a large gap between average and top earners.

Experience Indicators
Total Working Years shows a median around 10 years, with several outliers extending beyond 30 years.
Years at Company displays a heavily right-skewed distribution with numerous outliers beyond 20 years, while the majority (box) remains under 10 years.

Professional Development
Training Times Last Year shows a compact distribution with the box centered around 2-3 sessions, with a few outliers receiving up to 6 training sessions.

Work-Life Balance
The distribution is relatively symmetric across the 0-3 scale, with the box spanning levels 1-2, indicating moderate satisfaction with work-life balance across the workforce.

3.Density Plots

Key Insights and Findings

Demographic Distribution

Age Profile
The age density plot shows a unimodal distribution with peak density around 35-40 years. The curve exhibits a slight right skew, with a gradual tail extending to 60 years and a sharp drop-off below 25 years.

Educational Pattern
The education density reveals multiple peaks, with the highest at level 3, followed by significant peaks at levels 2 and 4. This multimodal distribution suggests distinct educational qualification clusters within the workforce.

Work-Related Metrics

Distance Distribution
The distance-from-home plot shows a rapidly declining curve with:

- Highest density at 0-5 units
- Secondary peak around 10 units
- Long tail extending to 30 units

Career Metrics
Total Working Years displays a right-skewed distribution peaking at 10 years.
Years at Company shows a sharp peak at 5-7 years with rapid decline thereafter.

Satisfaction and Development

Job Satisfaction
The distribution is notably bimodal with:

- Major peaks at levels 2 and 3

- Minor peaks at levels 0 and 1
- Similar density heights for the major peaks

Work-Life Balance
Shows a distinctive trimodal distribution:

- Dominant peak at level 2
- Secondary peak at level 1
- Smaller peak at level 3

Training and Compensation

Training Frequency
The training times distribution shows:

- Primary peak at 2-3 sessions
- Secondary peak at 3-4 sessions
- Multiple smaller modes beyond 4 sessions

Monthly Income
Exhibits a heavily right-skewed distribution with:

- Primary peak below median income
- Multiple small peaks in higher income ranges
- Long tail extending into higher income brackets

4.2 Bivariate Analysis: Analyzing Relationships and Dependencies

In this section, we explore the relationships between various pairs of variables in the dataset to uncover potential correlations and dependencies, which may help in predicting employee attrition.

Correlation Matrix

The **correlation matrix** shows the strength and direction of relationships between numerical variables. Values range from -1 to 1:

- A value **close to 1** indicates a strong positive correlation (as one variable increases, the other also increases).
- A value **close to -1** indicates a strong negative correlation (as one variable increases, the other decreases).
- A value **around 0** suggests no correlation between the variables.

For example, if there is a positive correlation between **JobSatisfaction** and **JobLevel**, this suggests that employees with higher satisfaction tend to hold higher job levels.

Scatter Plot

The **scatter plot** visualizes the relationship between two quantitative variables. By plotting data points on a Cartesian plane, it helps to identify trends, clusters, and outliers.

Example: In a **YearsAtCompany vs. MonthlyIncome** scatter plot, we may see that employees with longer tenures have a broader range of monthly incomes, while those with shorter tenures show more concentrated income values. This pattern could help to identify if employees are more likely to leave based on their income level and time at the company.

Pair Plot

A **pair plot** is a grid of scatter plots that visualizes pairwise relationships among multiple variables. It also includes histograms to show the distribution of each variable individually. This plot can help identify patterns or groupings of variables that may be correlated with employee attrition.

Example: A pair plot involving **Age**, **JobSatisfaction**, and **Exited** (attrition status) can reveal that employees who have exited (displayed in a different color) may cluster in specific areas of the plot, such as older individuals with low job satisfaction. This clustering could suggest that employees who are older and dissatisfied are more likely to leave the company.

By conducting this bivariate analysis, we aim to uncover key relationships between variables that are predictive of employee attrition, allowing for better decision-making in addressing retention strategies.

Corelation Matrix

Key Insights and Findings

Strong Positive Correlations

Experience and Income Metrics

- Total Working Years shows strong positive correlation with Monthly Income (0.77) and Age (0.68)
- Years at Company correlates significantly with Total Working Years (0.63) and Monthly Income (0.51)
- Age and Monthly Income display moderate positive correlation (0.50)

Weak or No Correlations

Job Satisfaction

- Shows remarkably weak correlations with all other variables (mostly around -0.01 to -0.02)
- Indicates job satisfaction is independent of factors like salary, age, or experience

Work-Life Balance

- Displays minimal correlation with all variables (coefficients near 0)
- Suggests work-life balance is maintained consistently across different employee segments

Distance From Home

- Shows negligible correlations with all variables
- Implies commute distance doesn't influence other workplace factors

Career Development

Training

- Training Times Last Year shows very weak negative correlations with most variables
- Suggests training opportunities are distributed independently of experience or position

Education

- Shows weak positive correlations with Age (0.21) and Total Working Years (0.15)
- Indicates educational level has minimal impact on other career metrics

Key Insights

- Career progression metrics (experience, income, age) are strongly interconnected
- Personal factors (job satisfaction, work-life balance) operate independently
- Training and education show surprising independence from career advancement metrics
- Location and distance factors have minimal impact on other variables

Scatter Plot

Key Insights and Findings

Age-Related Patterns

Career Progression

- Strong positive linear relationship between age and total working years, indicating natural career progression
- Age shows positive correlation with monthly income, though with high variance
- Years at company increases with age but shows significant scatter, suggesting varied retention patterns

Attrition Risk Factors

- Higher attrition (blue dots) appears more concentrated among younger employees (20-35 age range)
- Employees with longer distance from home show higher attrition across age groups
- Lower income brackets show higher concentration of attrition cases

Education and Training Insights

Educational Impact

- Education levels (1-5) are evenly distributed across age groups
- No clear relationship between education level and monthly income
- Higher education doesn't necessarily correlate with longer company tenure

Training Patterns

- Training frequency remains consistent across age groups
- No clear relationship between training and attrition

- Training distribution is similar regardless of distance from home

Job Satisfaction Analysis

Satisfaction Correlations

- Job satisfaction shows no clear correlation with age or income
- Work-life balance appears consistent across all age groups
- No strong relationship between satisfaction and years at company

Income Distribution

Salary Patterns

- Monthly income shows positive correlation with total working years
- Higher variance in income for employees with longer tenure
- Distance from home shows no clear impact on income levels

Critical Attrition Insights

High-Risk Groups

- Young employees with lower income
- Employees with longer commute distances
- Mid-career professionals with lower-than-expected income progression

Retention Factors

- Work-life balance appears consistent regardless of other factors
- Training opportunities distributed evenly across employee segments
- Job satisfaction varies independently of traditional career metrics

Key Recommendations

- Focus retention strategies on young employees and address commute-related concerns
- Maintain the positive aspects of work-life balance and training distribution

Pair Plot

Key Variable Relationships

The pairplot reveals several important patterns regarding IBM employee attrition:

Age and Experience
The scatter plots show that younger employees tend to have higher attrition rates, with the density of "Yes" responses being greater in the lower age ranges. Employees with more years at the company demonstrate lower attrition rates, suggesting that longer tenure correlates with higher retention.

Income Patterns
Monthly income displays a clear relationship with attrition. Lower-income employees show higher attrition rates, while the scatter plots indicate that employees with higher salaries are more likely to stay. This suggests compensation plays a crucial role in retention decisions.

Distance and Work-Life Balance
The distribution plots indicate that employees who live farther from work have increased attrition rates. Work-life balance scores show distinct patterns between those who stay and leave, highlighting its importance in retention.

Education and Job Satisfaction
The visualization reveals that job satisfaction has a notable impact on attrition, with lower satisfaction levels corresponding to higher attrition rates. Education levels show some variation in attrition patterns, though the relationship appears less pronounced than other factors.

Notable Correlations

- A clear correlation exists between years of experience and attrition, with newer employees showing a higher likelihood of leaving.
- The relationship between monthly income and years at the company appears positive, suggesting that longer-tenured employees generally earn more and are less likely to leave.

Recommendations
The data suggests focusing retention strategies on:

- Early-career employees
- Those with lower compensation levels
- Employees with longer commute distances
- Staff showing signs of job dissatisfaction

These insights can help develop targeted retention programs to reduce attrition rates effectively. trition rates effectively.

4.3 Multivariate Analysis

Multivariate analysis examines interactions among three or more variables, providing insights into complex relationships and patterns within the dataset.

Heatmap:
A heatmap visualizes correlations among multiple variables using color gradients to indicate the strength and direction of relationships.

Example: In the correlation heatmap for **Age**, **MonthlyIncome**, **YearsAtCompany**, and **ExitStatus**, a moderate positive correlation (0.29) exists between **Age** and **ExitStatus**, while a negative correlation (-0.30) is observed between **YearsAtCompany** and **MonthlyIncome**. These correlations help identify key factors related to employee retention and turnover.

Pairwise Relationships:
This technique analyzes relationships among multiple variables by assessing pairs together, revealing clusters and outliers.

Example: A pair plot for **Age**, **MonthlyIncome**, **YearsAtCompany**, and **ExitStatus** shows that employees who exited tend to cluster in specific areas, highlighting shared characteristics that influence job change behavior.

Principal Component Analysis (PCA):
PCA reduces dimensionality by transforming the dataset into uncorrelated variables known as principal components, facilitating visualization of data.

Example: Applying PCA to **Age**, **MonthlyIncome**, and **YearsAtCompany** reduces the data to two or three components, clearly distinguishing clusters of employees who exited from those who stayed.

Key Insights:
The analysis of multiple variables uncovers intricate relationships and interactions. For example, PCA helps illustrate how distinct groups of employees—those who exited versus those who stayed—can be differentiated based on their characteristics, thereby assisting in the identification of fundamental factors that influence employee retention and turnover.

1.Heat Map

Key Insights and Findings

Strong Positive Correlations
The correlation heatmap reveals several significant relationships:

Experience and Income

- Total Working Years and Monthly Income show a strong positive correlation (0.77)
- Years at Company and Total Working Years display a robust correlation (0.63)
- Monthly Income and Years at Company have a moderate positive correlation (0.51)

Age-Related Patterns

- Age and Total Working Years exhibit a strong positive correlation (0.68)
- Age and Monthly Income show a moderate positive correlation (0.50)
- Age and Years at Company have a weak positive correlation (0.31)

Weak or No Correlations
Work-Life Factors

- Work-Life Balance shows minimal correlation with other variables (all correlations < 0.05)
- Distance From Home has negligible correlations with most variables
- Job Satisfaction demonstrates very weak correlations with all other factors

Notable Insights
The heatmap suggests that:

- Career progression naturally links experience, age, and income
- Work-life balance operates independently of other employment factors
- Job satisfaction appears to be influenced by factors not captured in these variables
- Training times show minimal relationship with other employment metrics

These patterns can inform HR strategies by highlighting the independence of certain factors like **work-life balance** and **job satisfaction** from traditional career metrics.

2. Pair Plot

Key Insights and Findings

Age and Experience Relationships

Attrition by Age:

- Younger employees show higher attrition rates with denser clustering in the lower age ranges.
- Attrition probability decreases as age increases, showing a clear negative correlation.
- The distribution is right-skewed for both attrition groups.

Career Progression:

- **Total Working Years** and **Years at Company** show strong positive relationships.
- Employees with longer tenure demonstrate lower attrition rates.
- Experience metrics cluster differently between staying and leaving employees.

Income Patterns

Salary Distribution:

- **Monthly Income** shows a clear bimodal distribution.
- Higher attrition rates concentrate in lower income brackets.
- Income increases show positive correlation with retention.

Education Impact

- Higher education levels slightly correlate with increased income.
- Education level shows minimal direct impact on attrition.
- Distribution across education levels is relatively uniform.

Work-Life Factors

Distance and Satisfaction:

- **Distance from home** shows scattered distribution with no clear pattern.
- **Job satisfaction** levels cluster distinctly between attrition groups.
- **Work-life balance** scores show minimal correlation with other variables.

Training and Development:

- **Training times last year** shows discrete distribution.
- No strong correlation between training frequency and attrition.
- Training participation appears independent of other career metrics.

These insights suggest that retention strategies should focus on:

- **Early-career support** for younger employees, especially those in lower income brackets.
- **Competitive compensation** to improve retention, particularly for employees in lower income ranges.

- **Career development opportunities**, particularly for employees with longer tenure, to reduce attrition.

3. Principal Component Analysis

PCA reduces dataset dimensionality while preserving essential information, enabling visualization and clustering.

Findings:

Class Distribution: Overlap between attrition classes indicates that attrition factors are complex and not linearly separable.

Data Clustering: Central clustering shows that PCA dimensions capture general characteristics rather than specific attrition predictors.

Dimensionality Challenges: More features or higher-dimensional analysis is required for better distinction.

Practical Implications:

Employee attrition prediction requires sophisticated models beyond simple PCA, incorporating additional relevant variables.

Key Takeaways from Multivariate Analysis

Correlations reveal actionable patterns like the impact of age, income, and experience on retention.

Pair plots highlight clusters of employees likely to exit, emphasizing the need for strategic interventions in specific employee segments.

PCA insights underline the complexity of attrition, prompting advanced analytics for effective solutions.

By leveraging these insights, organizations can craft targeted strategies to improve employee retention and address turnover challenges effectively.

5 Regression Analysis

Regression and Classification Analysis of Employee Attrition

Simple Linear Regression

Insights:

R^2 Value: 0.022 (2.2% variance explained).

RMSE: 0.335 (moderate errors).

Key Observations:

Weak negative correlation: Employees with more working years are slightly less likely to leave.

Residuals show non-random patterns, suggesting a linear model may not be ideal.

Scatter plot highlights clustering due to the binary nature of the target variable (attrition = 0 or 1).

Conclusion:

Linear regression poorly fits the data, indicating the need for more robust models or alternative approaches like classification.

Multiple Linear Regression

Model Performance:

R^2 Value: 0.06 (6% variance explained, improvement over simple regression).

RMSE: 0.329.

Significant Predictors ($p < 0.05$):

Negative impact on attrition: Job satisfaction, work-life balance, age, training times last year.

Positive impact on attrition: Distance from home.

Diagnostics:

Non-normal residuals (Jarque-Bera = 573.732).

Minimal autocorrelation (Durbin-Watson = 2.104).

Conclusion:

Incorporating multiple factors improves the model's explanatory power slightly. However, the R^2 remains low, suggesting these features alone are insufficient for predicting attrition

Polynomial Regression

Performance Metrics:

R^2 Value: 0.0551 (5.51% variance explained).

RMSE: 0.3575.

Visual Insights:

The polynomial regression curve fits the data better than the linear model, capturing non-linear relationships.

Limitations:

Both linear and polynomial models struggle due to the binary nature of attrition.

Marginal improvement over linear regression.

Conclusion:

Polynomial regression captures some non-linear trends but still underperforms for binary classification problems.

Logistic Regression

Classification Metrics:

Accuracy: 87.07%.

Precision: 54.55% (moderate reliability in predicting attrition).

Recall: 39.34% (many attrition cases missed).

ROC-AUC: 0.8065 (good class separation).

Confusion Matrix Analysis:

True Negatives (360): Accurately predicted employees who stayed.

True Positives (24): Correctly identified some attrition cases.

False Negatives (37): Missed many actual attrition cases.

False Positives (20): Misclassified employees as leaving.

Model Strengths:

High overall accuracy and good discriminative ability.

Effective at predicting who will stay.

Model Limitations:

Low recall: The model struggles to identify actual attrition cases, potentially due to class imbalance.

Conclusion:

Logistic regression outperforms regression models for binary attrition prediction. However, its limitations in recall and handling class imbalance suggest a need for further optimization or exploring alternative classification methods.

Overview

For predicting employee attrition, classification models (e.g., logistic regression) are more suitable than regression approaches.

Address class imbalance through techniques like oversampling, undersampling, or using weighted loss functions.

Consider advanced models (e.g., random forests, gradient boosting) for improved performance.

Include additional features (e.g., engagement metrics, organizational culture) for better prediction accuracy.

6.Regularization Techniques

1. Ridge Regression

Performance: Exceptional accuracy with $R^2 \approx 1.0$, RMSE = 2.97e-05, and MAE = 2.16e-05.

Effectiveness: Handles multicollinearity efficiently via L2 penalty, ensuring robust predictive power.

Insight: May indicate potential overfitting; additional validation recommended.

2. LASSO Regression

Performance: Strong performance with $R^2 = 0.9278$, RMSE = 0.091, and MAE = 0.069.

Effectiveness: Performs feature selection through L1 penalty, sacrificing some accuracy for interpretability.

3. Elastic Net Regression

Performance: Balanced results with $R^2 = 0.9691$, RMSE = 0.060, and MAE = 0.045.

Effectiveness: Combines the strengths of Ridge and LASSO for handling multicollinearity and feature selection.

Conclusion:

Ridge Regression excels in predictive accuracy.

Elastic Net is the most versatile for balancing multicollinearity and feature importance.

Advanced Regression Techniques

1. Quantile Regression

Use Case: Predicts specific quantiles (e.g., median) of a dependent variable's distribution.

Application: Useful for skewed or non-normal data.

Significant Effects:

Positive: Age, Job Satisfaction, Monthly Income.

Negative: Distance from Home, Total Working Years, Training Times Last Year.

2. Poisson Regression

Use Case: Models count data with event occurrences over a fixed interval.

Key Insight: Assumes mean equals variance.

Significant Predictors: Age, Job Satisfaction, Work-Life Balance negatively associated; Distance from Home positively associated.

3. Negative Binomial Regression

Use Case: Suitable for overdispersed count data.

Effectiveness: Addresses variance exceeding the mean.

Key Predictors: Age, Education, Job Satisfaction, Training Times Last Year (negative); Monthly Income (marginally positive).

4. Zero-Inflated and Hurdle Regression

Use Case: Handles excessive zero counts in data.

Applications: Examples include zero-inflated Poisson for frequent zero counts in purchase data.

Significant Predictors: Age, Education, Job Satisfaction, and others improve fit by addressing excess zeros.

5. Cox Regression (Survival Analysis)

Use Case: Models time-to-event data, predicting hazard rates.

Performance: Concordance = 0.79 (strong predictive ability).

Insight: Used for medical research or time-to-failure modeling.

6. Partial Least Squares Regression (PLSR)

Use Case: Handles high-dimensional datasets where predictors exceed observations.

Effectiveness: Captures relationships in predictors and responses through latent variables.

7. Principal Component Regression (PCR)

Use Case: Reduces dimensionality via PCA before applying linear regression.

Insight: Reduces multicollinearity; results may align closely with PLSR.

Conclusion on Advanced Regression Techniques

Key Predictors

Age, Job Satisfaction, Work-Life Balance, and Training Times Last Year are the main predictors consistently influencing attrition across various models.

Focus should be placed on improving job satisfaction, enhancing work-life balance, and increasing training opportunities to minimize attrition.

Zero-Inflated Poisson Insights

The Zero-Inflated Poisson model identifies a significant number of employees who are highly unlikely to leave the company, suggesting the need for tailored retention strategies for these groups.

Dimensionality Reduction (PLSR & PCR)

These techniques helped address multicollinearity and reveal complex interactions among predictors, offering robust insights for attrition analysis, especially when dealing with correlated variables.

Practical Implications for IBM

Develop targeted retention programs for employees, especially focusing on specific age or tenure groups.

Use advanced techniques like Zero-Inflated Poisson and PLSR/PCR to gain deeper insights into attrition patterns.

Continuously monitor and address predictors like distance to work, job satisfaction, and income to improve retention.

7.Selected Model Evaluation and Validation

Evaluation Metrics for All Models

Regression Models:

Simple Linear Regression:

R^2: 0.0218, RMSE: 0.3355

Multiple Linear Regression:

R^2: 0.0602, RMSE: 0.3288

Polynomial Regression:

R^2: 0.0551, RMSE: 0.3575, MAE: 0.2556

Classification Model:

Logistic Regression:

Accuracy: 87.07%, Precision: 0.5455, Recall: 0.3934, ROC-AUC: 0.8065

Comparison of Model Performance

Regression Models

Multiple Linear Regression: Slightly better performance than simple linear regression with a higher R^2 (0.0602 vs. 0.0218) and lower RMSE (0.3288 vs. 0.3355).

Polynomial Regression: Performs worse than both linear models with lower R^2 (0.0551) and higher RMSE (0.3575).

Classification Model

Logistic Regression excels with high accuracy (87.07%) and a solid ROC-AUC score (0.8065), showing strong discriminative power. However, the relatively lower precision (0.5455) and recall (0.3934) indicate some challenges in identifying true attrition cases.

Summary of Findings

All regression models exhibit poor performance, explaining very little of the variance in employee attrition.

Logistic Regression performs significantly better for predicting attrition as a classification task.

Precision and recall scores highlight a trade-off in identifying attrition cases.

The poor performance of regression models suggests that the relationship between variables and attrition is likely non-linear.

Insights and Implications

Complex Attrition Factors: Employee attrition is influenced by factors not captured by the current set of variables.

Classification Over Regression: Logistic regression is better suited for predicting attrition, though improvements in feature selection and balancing are needed.

Feature Engineering: Incorporating more relevant features or interactions could improve predictive power.

Model Interpretability and Further Improvements

Interpretability Techniques

Intrinsic Analysis: Start with interpretable models like linear regression or decision trees for basic insights.

Post Hoc Analysis: Use techniques like LIME or SHAP to explain individual predictions.

Global Surrogate Models: Train simpler models to approximate complex models.

Feature Importance: Use permutation importance to identify key features.

Partial Dependence Plots: Visualize feature-target relationships.

Advanced Techniques

Feature Engineering: Consider polynomial or interaction terms to capture complex relationships.

Domain Knowledge: Integrate domain-specific knowledge to create relevant features.

Ensemble Methods: Explore ensemble techniques or deep learning approaches for better performance while balancing interpretability.

Customer Behavior and Attrition Prediction

Segmentation Strategies

Demographic Segmentation: Group employees by age, gender, total working years, and income to analyze attrition risks.

Behavioral Segmentation: Categorize employees based on engagement levels, performance, and interaction patterns.

Lifecycle Segmentation: Analyze employees by their job level and tenure to target retention efforts based on different lifecycle stages.

Feedback and Satisfaction Segmentation: Use satisfaction surveys or performance reviews to track satisfaction levels and predict attrition.

Implications

Using segmentation strategies allows for more personalized and effective interventions by identifying specific risk factors for different employee groups. This enables IBM to tailor retention efforts to meet the unique needs of various segments, improving overall employee retention.

8.Conclusion

This project successfully demonstrated the application of logistic regression combined with regularization techniques to predict employee attrition. The inclusion of regularization helped mitigate overfitting, enhancing the model's generalizability on unseen data. The insights derived from the model are valuable for organizations looking to reduce attrition and improve their employee retention strategies.

The analysis revealed that key employee attributes, including age, job satisfaction, monthly income, distance from home, work-life balance, and performance rating, have a significant relationship with employee attrition. The results provide strong evidence to reject the null hypothesis, confirming that these factors play a crucial role in influencing whether employees stay or leave the organization.

Further improvements to the model could involve:

Incorporating additional features to capture more complex patterns of attrition.

Exploring advanced modeling techniques, such as decision trees or ensemble methods, to improve prediction accuracy.

Refining customer segmentation strategies, which could allow businesses to tailor interventions more effectively and make more informed decisions.

Ultimately, this approach offers a foundation for businesses to understand attrition dynamics better and design more effective retention programs.

References

DataCamp. "An Introduction to Exploratory Data Analysis." Available at: https://www.datacamp.com/community/tutorials/exploratory-data-analysis-python

Seaborn Documentation. "Seaborn Overview." Available at: https://seaborn.pydata.org/

Matplotlib Documentation. "Matplotlib User Guide." Available at: https://matplotlib.org/stable/users/index.html

Medium. "A Comprehensive Guide to Exploratory Data Analysis (EDA)." Available at: https://medium.com/

Kaggle. "Customer Churn Analysis with Python." Available at: https://www.kaggle.com/learn/customer-churn-analysis-with-python

Towards Data Science. "Exploratory Data Analysis (EDA) Visualization Using Pandas." Available at: https://towardsdatascience.com/