A black background with a black square

Description automatically generated with medium confidence

***IBM HR Analytics Employee Attrition & Performance***

***Course Name:*** *BCIS 5110 Section 001 - Programming Languages for Business Analytics (Fall 2024 1)*

***Instructor Name:*** *Ling Ge*

***Analysis By:*** *Rakshitha Chintalapudi*

**Executive Summary:**

Employee attrition presents significant difficulties for organizations, leading to substantial costs, decreased productivity, and disruptions in team dynamics. This study, titled **"IBM HR Analytics Employee Attrition & Performance,"** seeks to uncover the primary factors contributing to employee turnover and provide actionable approaches to improve retention. Leveraging a dataset of 1,470 employees, which contains details on demographics, job roles, compensation, work conditions, and attrition outcomes, the analysis integrates both descriptive and predictive methods to deliver meaningful findings.

The descriptive analysis highlighted critical trends in attrition. Departments such as Sales and Human Resources exhibit the highest turnover rates, while younger employees and those in lower-compensated roles, such as Sales Representatives, are more likely to leave. Gender-based attrition shows slightly higher rates for males, while employees with low job satisfaction are at greater risk of leaving. These findings highlight the importance of tackling department-specific issues, improving job satisfaction, and developing retention strategies tailored to employees at higher risk of leaving.

A predictive model, developed using **XGBoost** and balanced with **SMOTE**, achieved strong performance metrics, including 91% accuracy and a ROC-AUC score of 0.97. Key predictors of attrition include overtime, marital status, frequent business travel, and short tenure. These findings enable HR teams to proactively recognize and resolve employees at risk of leaving. By improving work-life balance, offering growth opportunities, and addressing role-specific issues, organizations can leverage this analysis to reduce turnover, foster employee satisfaction, and ensure workforce stability.

**Project Motivation Background**

Employee turnover, also known as attrition, poses substantial challenges for organizations globally, often resulting in higher costs for recruitment and training, diminished team morale, and disruptions in productivity. Understanding the factors driving employee departures is critical to building a stable and engaged workforce.

This project, titled **IBM HR Analytics Employee Attrition & Performance**, seeks to address this issue by examining patterns and predictors of employee turnover. By leveraging data analytics, the project aims to provide actionable findings to help organizations reduce attrition rates and improve employee satisfaction.

The dataset, sourced from IBM HR Analytics and publicly available on Kaggle, contains 1,470 employee records across 35 variables. These variables encompass demographic data (e.g., age, gender), job-related attributes (e.g., role, department, salary), and work environment metrics (e.g., job satisfaction, overtime). The focus of the study is to analyze the relationships between these factors and employee attrition, as well as to develop a predictive model to identify employees who may be at risk of leaving.

The research addresses the following key questions:

* **Descriptive Questions:**
  1. What is the average monthly income across different job roles?
  2. Which department has the highest attrition rate?
  3. How does attrition vary by gender?
  4. What is the relationship between age and the likelihood of attrition?
  5. What is the distribution of job satisfaction levels among employees who left versus those who stayed?
* **Predictive Question:** Can we determine the likelihood of an employee leaving the organization based on their demographic characteristics, job role, and workplace conditions?

By addressing these questions, this project provides insight on the key factors influencing attrition and equips organizations with the tools to tackle it proactively. Utilizing both descriptive and predictive analytics, the study offers actionable findings to develop data-driven strategies focused on improving employee retention.

**Data Description**

This project utilizes the IBM HR Analytics Employee Attrition & Performance dataset, obtained from Kaggle. The dataset contains detailed information about 1,470 employees, encompassing their demographics, job roles, and work environment. Its structure supports both descriptive and predictive analyses to explore employee attrition.

**Dataset Link**:  
<https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?resource=download>

**Dataset Overview**

* **Number of Observations:** 1,470 employees
* **Number of Variables:** The dataset contains **35 features**, which include:
* **Demographic Information:** Attributes like **Age**, **Gender**, and **Marital Status**.
* **Job Details:** Information on **Job Role**, **Department**, **Job Level**, and **Years at Company**.
* **Compensation:** Details such as **Monthly Income**, **Stock Option Level**, and **Percent Salary Hike**.
* **Work Environment:** Factors like **Job Satisfaction**, **Work-Life Balance**, and **Overtime**.
* **Performance Metrics:** Includes **Performance Rating** and **Training Times Last Year**.
* **Target Variable:** **Attrition** (indicates whether the employee left the company: **Yes** or **No**).

**Data Types**

* **Numerical Variables:** Continuous and discrete values, e.g., Age, Monthly Income, Years at Company.
* **Categorical Variables:** Binary and multi-class categories, e.g., Gender, Job Role, Attrition.

**Target Variable**

* The dependent variable, Attrition, indicates whether an employee left the company.
  + **Value = Yes (1):** Employee left.
  + **Value = No (0):** Employee stayed.

**Data Quality**

The dataset is clean and complete, with no missing values. However, the following preparations were necessary:

* **Encoding Categorical Variables:** Binary categories like Attrition and Overtime were converted to 0/1, and multi-class variables like Job Role were one-hot encoded.
* **Outliers:** Outliers in numerical columns like Monthly Income were handled using the IQR method to avoid skewing the results.
* **Irrelevant Columns:** Constant or non-informative columns, such as Over18 and EmployeeNumber, were removed.

This dataset provides a solid foundation for analyzing employee attrition, enabling insights into key drivers and predictors of turnover.

**Data Preparation**

Data preparation is a crucial step in any analysis, ensuring the dataset is clean, consistent, and ready for modeling. For this project, the following preprocessing steps were performed on the IBM HR Analytics Employee Attrition & Performance dataset:

**1. Dropping Irrelevant Columns**

Certain columns were identified as irrelevant or uninformative for the analysis and were removed:

* EmployeeCount and StandardHours: These columns had constant values for all employees.
* Over18: Provided no meaningful variation (all employees were over 18).
* EmployeeNumber: A unique identifier that did not contribute to analysis.

**2. Encoding Categorical Variables**

To make the dataset suitable for machine learning models, categorical variables were encoded:

* **Binary Variables:**
  + Attrition and OverTime were converted to binary values (Yes = 1, No = 0).
* **Multi-Class Variables:**
  + Variables like JobRole, BusinessTravel, and Department were one-hot encoded to represent categories as separate columns.

**3. Handling Outliers**

Outliers in numerical columns, particularly MonthlyIncome, were addressed using the Interquartile Range (IQR) method to ensure they did not skew the analysis.

**4. Splitting Data for Modeling**

The dataset was split into training and testing sets to evaluate the predictive model:

* **Training Data:** 70% of the data for model training.
* **Testing Data:** 30% of the data for evaluating model performance.

**Summary of Prepared Dataset**

After preprocessing, we have 44 columns, including both numerical and encoded categorical variables. The target variable, Attrition, remained the focal point for the predictive model.

These preprocessing steps ensured the dataset was clean, structured, and ready for both exploratory analysis and predictive modeling.

**Exploratory Data Analysis**

The exploratory data analysis (EDA) focuses on uncovering key patterns and trends in employee attrition. This section answers the five descriptive questions using statistical summaries and visualizations.

**Question 1: What is the average monthly income across different job roles?**

* **Findings:**
  + Managers earn the highest average monthly income ($17,181.68), followed by Research Directors ($16,033.55).
  + Sales Representatives have the lowest average income ($2,626.00), which may contribute to higher turnover in this role.
* **Visualization:** A bar chart highlights income disparities across job roles, emphasizing the wide variation in salaries.

**Average Monthly Income by Job Role:**

A graph of a number of people

Description automatically generated with medium confidence

**Question 2: Which department has the highest attrition rate?**

* **Findings:**
  + The Sales department has the highest attrition rate (20.63%), followed by Human Resources (19.05%).
  + Research & Development exhibits the lowest attrition rate (13.84%).
* **Visualization:** A bar chart visually compares attrition rates across departments, with Sales standing out as the most affected.

**Attrition Rate by Department:**

A graph with different colored bars

Description automatically generated

**Question 3: How does attrition vary by gender?**

* **Findings:**
  + Male employees exhibit a slightly higher attrition rate (17.01%) compared to female employees (14.80%).
* **Visualization:** A bar chart shows the proportional difference in attrition rates between genders.

**Attrition Rate by Gender:**

**A graph with a number of people

Description automatically generated with medium confidence**

**Question 4: What is the relationship between age and the likelihood of attrition?**

* **Findings:**
  + Younger employees (under 35 years) are more likely to leave, as indicated by a higher density of attrition in this age group.
  + Employees aged 40 years and above show higher retention rates.
* **Visualization:** A KDE plot and a boxplot illustrate the relationship between age and attrition, highlighting the concentration of attrition among younger employees.

**Visualize the relationship between age and attrition using a boxplot**:

**A graph of a number of people

Description automatically generated with medium confidence**

**Visualize the age distributions using kdeplot:**

**A graph of a normal distribution

Description automatically generated**

**Question 5: What is the distribution of job satisfaction levels among employees who left versus those who stayed?**

* **Findings:**
  + Employees who left are more concentrated in lower satisfaction levels (1 and 2), while those who stayed are distributed more evenly, with a larger proportion in higher satisfaction levels (3 and 4).
* **Visualization:** A stacked bar chart compares job satisfaction distributions for employees who stayed versus those who left.

**Key Insights from EDA**

1. Departments like Sales and roles like Sales Representatives face the highest turnover, indicating a need for targeted interventions.
2. Younger employees and those with low job satisfaction are at a higher risk of leaving, highlighting potential areas for improvement in engagement and retention strategies.
3. Income disparities between roles may contribute to differences in attrition rates, emphasizing the need for fair compensation practices.

**Job Satisfaction Distribution:**

A graph of different colored squares

Description automatically generated

**Predictive Question:** Can we determine the likelihood of an employee leaving the organization based on their demographic characteristics, job role, and workplace conditions?

The predictive question examines the feasibility of building a model to forecast employee attrition based on key predictors identified during the exploratory data analysis (EDA). Insights from the EDA provided a solid foundation for developing the predictive model:

* **Key Predictors Identified:**
  + **Overtime:** Employees working overtime are significantly more likely to leave.
  + **Marital Status:** Single employees exhibit higher attrition rates compared to married employees.
  + **Frequent Business Travel:** Employees with frequent travel responsibilities are more prone to attrition.
  + **Years at Company:** Employees with shorter tenure have a higher likelihood of leaving.

These variables, along with others, were used to train an **XGBoost model**, which achieved the following performance metrics:

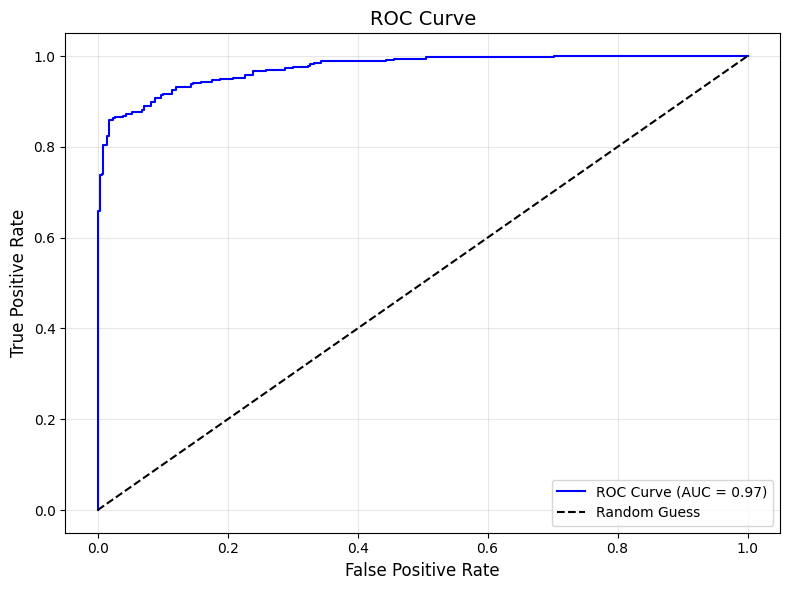
* **Accuracy:** 91%
* **ROC-AUC Score:** 0.97

**Precision-Recall Curve:**

A graph of a curve

Description automatically generated

**Roc Curve:**

****

These results highlight the model's strong ability to predict employee attrition while balancing precision and recall effectively. The insights derived from the EDA ensured that the model leveraged meaningful trends and relationships in the data, enabling HR teams to make data-driven decisions for improving retention.

**Models and Analysis**

This section is dedicated to developing a predictive model to address the question: Can employee attrition be predicted based on their demographics, job role, and work environment? Logistic regression was selected due to its simplicity, interpretability, and effectiveness in binary classification problems.

**Model Development**

* **Target Variable:**  
  The dependent variable is Attrition, where:
  + 1 indicates the employee left the company.
  + 0 indicates the employee stayed.
* **Independent Variables:**  
  Based on insights from EDA, key predictors included:
  + **Job Satisfaction:** Employees with lower satisfaction levels are more likely to leave.
  + **Overtime:** Employees working overtime show higher attrition rates.
  + **Monthly Income:** Lower-income employees are more prone to attrition.
  + **Age:** Younger employees are more likely to leave.
  + Other variables: Department, Job Role, YearsAtCompany, WorkLifeBalance.
* **Preprocessing Steps:**
  + Categorical variables were encoded (e.g., one-hot encoding for Job Role and Department).
  + Numerical variables were standardized for consistent scaling.
  + Outliers in variables like Monthly Income were handled using the IQR method.
* **Training and Testing:** The dataset was split into:
  + **70% training data** for model development.
  + **30% testing data** for evaluation.

**Model Performance**

The XGBoost model was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, providing a comprehensive assessment of its performance.

* **Classification Report:**

Precision Recall F1-Score Support

Stayed 89% 93% 91% 340

Left 92% 88% 90% 335

The model demonstrates balanced performance, with high precision (92%) and recall (88%) for predicting employees who left. It also performs well in identifying employees who stayed, with an F1-Score of 91%.

* Confusion Matrix **(Threshold = 0.4):**

Stayed Left

Stayed 316 24

Left 40 295

**Confusion Matrix Heatmap:**

A diagram of a heatmap

Description automatically generated

The model correctly identifies **316 employees who stayed** and **295 employees who left**, while misclassifying **40 employees who left as stayed** and **24 employees who stayed as left**.

* **ROC-AUC Score:** The model achieved an **AUC score of 0.97**, indicating excellent discriminatory ability in distinguishing between employees who stayed and those who left.

**Key Insights**

1. **Attrition Trends:** Attrition is highest among younger employees, with departments like Sales and Human Resources experiencing the highest turnover rates. Roles with lower compensation, such as Sales Representatives, and employees with lower job satisfaction are particularly at risk.
2. **Key Predictors:** Overtime, marital status (single employees), frequent business travel, and short tenure are the most influential factors driving attrition. Employees working overtime or those with fewer years at the company are significantly more likely to leave.
3. **Model Insights:** The XGBoost model achieved a **91% accuracy** and an **AUC score of 0.97**, effectively identifying employees at risk of leaving. This allows HR teams to proactively address retention challenges and implement targeted interventions.

**Findings and Managerial Implications**

This section summarizes the insights derived from the descriptive and predictive analyses and provides actionable recommendations for addressing employee attrition.

**Key Findings**

1. **Income Disparity and Attrition:**
   * Employees in lower-paying roles, such as Sales Representatives, are more likely to leave the organization.
   * Managers and Research Directors, who have higher average incomes, exhibit better retention rates.
2. **Departmental Attrition Trends:**
   * Sales has the highest attrition rate (20.63%), followed by Human Resources (19.05%).
   * Research & Development shows the lowest attrition rate (13.84%), indicating greater stability in this department.
3. **Gender and Attrition:**
   * Male employees have a slightly higher attrition rate (17.01%) compared to female employees (14.80%).
4. **Age and Attrition:**
   * Younger employees (under 35 years) exhibit higher attrition rates, while employees aged 40 and above are more likely to stay.
5. **Job Satisfaction and Attrition:**
   * Employees with lower job satisfaction levels (1 and 2) are significantly more likely to leave compared to those with higher satisfaction levels (3 and 4).
6. **Predictive Insights from the Logistic Regression Model:**
   * **Strong Performance Metrics:** The logistic regression model achieved an **accuracy of 91%** and a **ROC-AUC score of 0.97**, demonstrating its effectiveness in distinguishing between employees who stay and those likely to leave.
   * **Key Predictors of Attrition:** The most influential factors driving attrition include **overtime**, **marital status (single employees)**, **frequent business travel**, and **short tenure**, enabling targeted interventions to reduce turnover.

**Managerial Implications**

1. **Targeted Retention Strategies:**
   * **Departments:** Focus retention efforts on departments like Sales and Human Resources, where attrition rates are highest. Address role-specific challenges and workloads to improve retention.
   * **Demographics:** Younger employees and those with low job satisfaction should be prioritized for engagement programs and mentorship opportunities.
2. **Compensation and Benefits:**
   * Reassess compensation structures for lower-paying roles, such as Sales Representatives, to align with industry standards and employee expectations.
3. **Job Satisfaction Improvements:**
   * Conduct regular surveys to assess job satisfaction levels and address areas of concern.
   * Implement initiatives such as flexible work policies, professional development programs, and opportunities for career growth to improve employee satisfaction.
4. **Proactive Use of Predictive Analytics:**
   * Leverage the logistic regression model to identify at-risk employees early and intervene with tailored solutions.
   * Continuously refine the model by incorporating additional data, such as performance reviews and employee feedback, to improve prediction accuracy.
5. **Work-Life Balance:**
   * Reduce overtime workloads, as employees working overtime exhibit higher attrition rates. Encourage a healthy work-life balance through policy changes and supportive management practices.

**Conclusion**

The findings from this analysis provide a clear roadmap for addressing employee attrition through data-driven decision-making. By focusing on high-attrition areas and implementing targeted strategies, organizations can reduce turnover, improve employee satisfaction, and enhance overall productivity.

**References**

* **IBM HR Analytics Employee Attrition & Performance Dataset (Kaggle):**  
  <https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>
* **Scikit-learn: Machine Learning in Python:**  
  https://scikit-learn.org/stable/
* **Matplotlib: Visualization with Python:**  
  <https://matplotlib.org/>
* **Pandas: Python Data Analysis Library:**  
  <https://pandas.pydata.org/>
* **NumPy: The Fundamental Package for Scientific Computing with Python**  
  <https://numpy.org/>
* **XGBoost Documentation: Scalable, Portable, and Distributed Gradient Boosting Library.**  
  <https://xgboost.readthedocs.io/>