Date:20/07/24

ACME Attrition BI Report

Name:Krishna Priya S

Abstract:

Predicting employee attrition can help organizations take the necessary steps to retain talent well within time. Based on our analysis, attrition rates were higher in younger employees, doing overtime, having lower monthly incomes and working for a shorter period of time.

Introduction:

Employee attrition refers to an employee's voluntary or involuntary resignation from a workforce. Organiza□tions spend many resources in hiring talented employees and training them. Every employee is critical to a company's success. Our goal is to predict employee attrition and identify the factors contributing to an employee leav□ing a workforce. We discuss various classification models on our dataset and assess their performance using different metrics such as accuracy, precision, recall and F1 score. We also analyze the dataset to identify key factors contributing to an employee leaving a workforce. Our project will assist organizations in gaining fresh insights into what drives attrition and thus enhance retention rate.

Tools used : PowerBI

Python Code: VS Code

Dataset :

Dataset Review We used is the IBM Employee Attrition dataset from Kaggle. It contains 35 columns and 1470 rows and has a mix of numerical and categorical features.

```
{'Age': 44,
 'BusinessTravel': 0,
 'DailyRate': 489,
 'Department': 1,
 'DistanceFromHome': 23,
 'Education': 3,
 'EducationField': 3,
 'EnvironmentSatisfaction': 2,
 'Gender': 1,
 'HourlyRate': 67,
 'JobInvolvement': 3,
 'JobLevel': 2,
 'JobRole': 2,
 'JobSatisfaction': 2,
 'MaritalStatus': 1,
 'MonthlyIncome': 2042,
 'MonthlyRate': 25043,
 'NumCompaniesWorked': 4,
 'OverTime': 0,
 'PercentSalaryHike': 12,
 'PerformanceRating': 3,
 'RelationshipSatisfaction': 3,
 'StockOptionLevel': 1,
 'TotalWorkingYears': 17,
 'TrainingTimesLastYear': 3,
 'WorkLifeBalance': 4,
 'YearsAtCompany': 3,
 'YearsInCurrentRole': 2,
 'YearsSinceLastPromotion': 1,
 'YearsWithCurrManager': 2}
```


Code:

Data cleaning and preprocessing, including handling missing values, removing duplicates, and treating outliers.Conducting exploratory data analysis (EDA) to understand relationships between different features and the target variable (Attrition).Performing feature engineering by encoding categorical variables and scaling numerical features.Training and evaluating machine learning models, specifically Random Forest Classifier and Gradient Boosting Classifier, to predict employee attrition.Conducting K-Fold cross-validation to assess the models' performance.Performing hyperparameter tuning using GridSearchCV to optimize the models' performance.Analyzing feature importance to identify the most influential factors contributing to employee attrition.

Attached code in repository for reference.

Description of Processing:

Data Preparation:

1. Importing Libraries:

- Libraries such as `pandas`, `numpy`, `matplotlib`, `seaborn`, and several from `sklearn` are imported for data processing, visualization, and model building.

2. Loading the Dataset:

- The employee data is loaded using `pandas.read_csv`.

3. Data Cleaning:

- Handling Missing Values: The mean of the 'Age' column is calculated and used to fill missing values.

- Removing NaNs and Duplicates: Rows with NaN values are dropped, and duplicate rows are removed.

- Outlier Detection and Treatment: Outliers in the 'MonthlyIncome' column are removed using the Interquartile Range (IQR) method.

Exploratory Data Analysis (EDA)

1. Histograms:

- Histograms are plotted for a univariate analysis of the dataset.

2. Boxplot:

- A boxplot is created to show the impact of job satisfaction on attrition.

3. Distribution Plots:

- Distribution plots for the 'Age' column and count plots for attrition by department are created.

4. Descriptive Analytics:

- A pairplot of key attrition-related features is plotted to understand relationships between them.

5. Correlation Analysis:

- A heatmap of the correlation matrix is plotted to visualize the relationships between different numerical features.

Statistical Analysis:

1. T-test:

- A t-test is performed to compare 'MonthlyIncome' between employees with and without attrition. The t-statistic and p-value are calculated to determine the significance of the difference.

Feature Engineering:

1. Encoding Categorical Variables:

- Categorical variables are encoded using `LabelEncoder`.

2. Feature Scaling:

- Numerical features are scaled using `StandardScaler`.

Model Building

1. Data Splitting:

- The data is split into training and testing sets using `train_test_split`.

2. Model Training and Evaluation:

- Two models, `RandomForestClassifier` and `GradientBoostingClassifier`, are trained and evaluated using accuracy, precision, recall, confusion matrix, and classification report.

3. K-Fold Cross Validation:

- Cross-validation is performed on both models to assess their performance across different folds of the data.

4. <u>Hyperparameter Tuning:</u> - Hyperparameter tuning is done using `GridSearchCV` to find the best parameters for both Random Forest and Gradient Boosting models.

5. <u>Feature Importance:</u>

- The importance of each feature is determined using the trained Random Forest model and visualized using a bar plot.

<u>ANALYSIS:</u>

The analysis reveals that attrition predominantly affects younger employees, particularly those aged 29 and 32. These age groups show higher attrition rates compared to older employees. This suggests that younger employees may be more likely to leave the company, potentially due to seeking better opportunities, career advancement, or dissatisfaction with their current roles and compensation. This insight highlights the need for targeted retention strategies focusing on younger employees to reduce attrition rates**.**

<u>OUTPUT:</u>

```
Random Forest K-Fold Cross Validation:
Accuracy:  0.8451296939439981
Precision:  0.7355128205128204
Recall:  0.14662349676225714
F1 Score:  0.24233207470495605
```

```
    # K-Fold Cross Validation on Gradient B
    print("\nGradient Boosting K-Fold Cross
    cross_validation(gb_clf, X, y)
```

```
Gradient Boosting K-Fold Cross Validation:
Accuracy:  0.8562079444323855
Precision:  0.6799552299552301
Recall:  0.3109158186864015
F1 Score:  0.4202358412703946
```

Integration with the PowerBI Dashboard

The insights and findings from this code can be effectively integrated into a PowerBI dashboard to provide a comprehensive visualization and reporting solution for employee attrition analysis. The dashboard can include the following components:

Attrition: A summary of the overall attrition rate, including the percentage of employees who have left the company.

Attrition by Department: A bar chart or stacked bar chart showing the distribution of attrition across different departments.

Attrition by Job Role: A visualization of attrition rates for various job roles within the company.

Demographic Analysis: Charts showing attrition patterns based on age, gender, and education level.

Key Factors Influencing Attrition: A visualization of the most important features identified by the machine learning models.

Predictive Insights: A section showcasing the predictions made by the machine learning models and their accuracy.

Employee Satisfaction Metrics: Visualizations of factors like job satisfaction, work-life balance, and job involvement, which were identified as important predictors of attrition.

Detailed Description of Charts:

1. No of Employee Count by Gender:

This bar chart shows the distribution of employees by gender (Female and Male) across different age groups.

Insights:

The majority of employees fall within the 25-34 age range. Females slightly outnumber males in this category. The number of employees decreases significantly in the under 25 and over 55 age groups.

2. Attrition Count by EducationField:

This donut chart displays the distribution of attrition across various education fields.

Insights:

Life Sciences has the highest attrition rate (37.5%), followed by Medical (26.58%) and Marketing (14.77%). This indicates a higher turnover in these fields.

3. Attrition by Education:

- This horizontal bar chart illustrates the attrition count by different education levels.

-Insights:

Employees with a Bachelor's degree show the highest attrition, suggesting that individuals with undergraduate degrees are more likely to leave the company.

4. Attrition by Department:

This pie chart represents the attrition distribution across different departments.

Insights:

The Sales department has the highest attrition rate (42.4%), followed by Research & Development (30.5%) and Human Resources (27.1%). Targeted retention strategies might be necessary for the Sales department.

5. Sum of Percent Salary Hike by Age:

-This bar chart shows the total percentage of salary hikes received by employees, grouped by age.

- Insights:

Employees aged 25-34 received the highest total salary hikes, indicating a focus on rewarding younger employees to retain them.

6. Attrition Count by Years at Company:

This line chart shows the number of attritions relative to the number of years employees have been with the company.

- Insights:

There is a peak in attrition for employees who have been with the company for less than one year. The rate decreases significantly and remains low for employees with longer tenure.

7. Attrition Count by Age:

This bar chart presents the number of attritions by different age groups.

- Insights: Employees aged 29 and 32 have the highest attrition rates, suggesting that individuals in their late 20s and early 30s are more likely to leave the company.

8. Job Role:

This table shows the number of employees by job role across different departments.

- Insights:

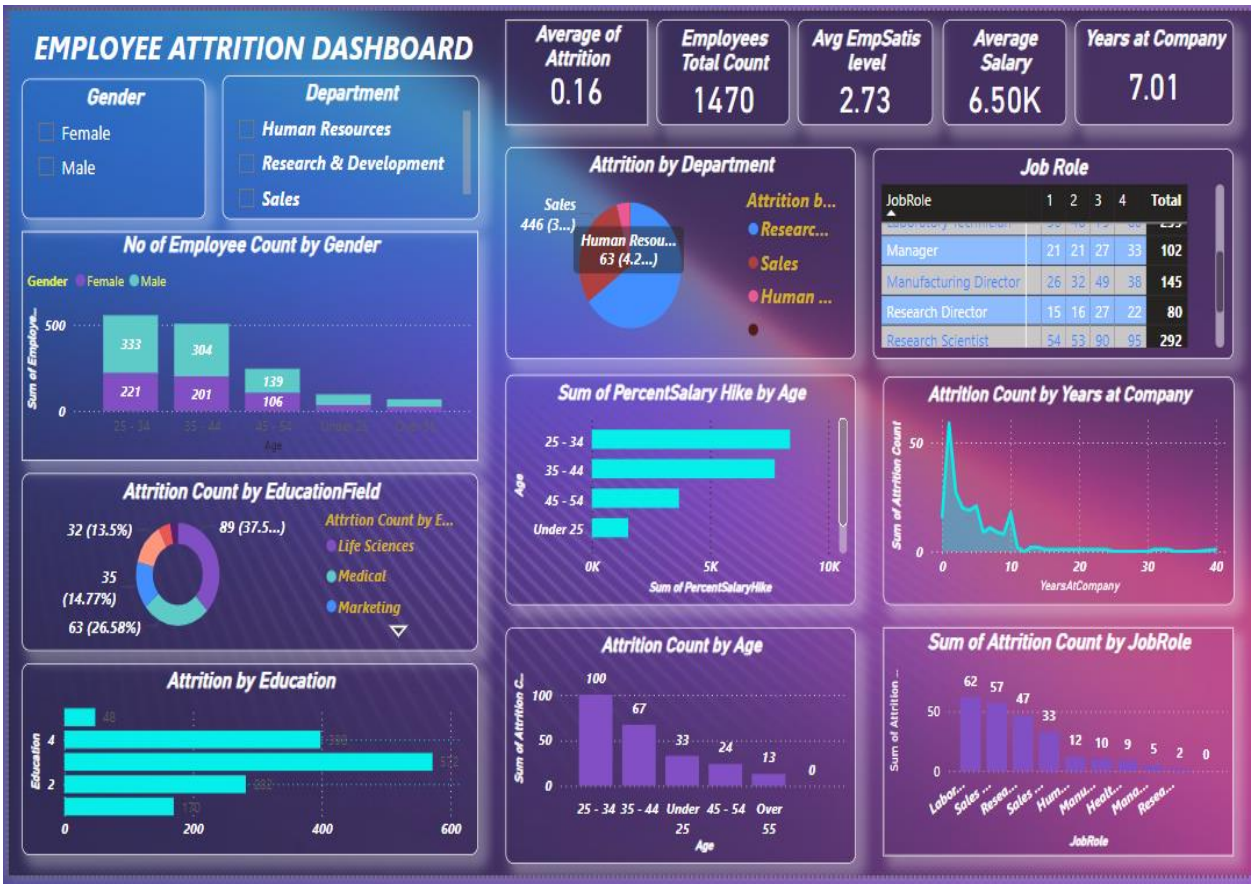
The highest number of employees are in the Research Scientist role (292), followed by Laboratory Technician (259), indicating a concentration of employees in technical roles.

9. Sum of Attrition Count by JobRole:

- This bar chart shows the attrition count by job role.

- Insights: Sales Executives (62) and Laboratory Technicians (57) have the highest attrition counts, highlighting the need for targeted retention strategies in these roles.

DASHBOARD:



Conclusion:

The dashboard and the Python code provide a comprehensive analysis of employee attrition at ACME. The dashboard offers visual insights into various factors affecting attrition, while the Python code performs detailed data cleaning, EDA, statistical analysis, feature engineering, and model building to predict employee attrition.The analysis of ACME's employee attrition data reveals several critical insights. Age plays a significant role in turnover, with employees in their late 20s and early 30s, particularly those aged 29 and 32, showing the highest attrition rates. This indicates potential challenges for younger employees, possibly related to career development or job satisfaction.

Departmental analysis highlights that the Sales department faces the highest attrition rate, followed by Research & Development and Human Resources. This suggests a need for targeted retention strategies in these areas. Specific job roles, such as Sales Executives and Laboratory Technicians, also exhibit higher attrition, pointing to job stress or dissatisfaction in these positions.Education level and field also influence attrition, with employees holding a Bachelor's degree and those from Life Sciences, Medical, and Marketing fields more likely to leave. This may reflect broader career opportunities or unmet expectations within the company. Compensation emerges as a crucial factor, with significant differences in monthly income between those who stay and those who leave, underscoring the importance of competitive salary packages.Job satisfaction and work-life balance are pivotal in predicting attrition. Employees with lower satisfaction and poor work-life balance are more inclined to leave, indicating that improvements in these areas could enhance retention.