Problem Statement: Predicting Employee Attrition

Dataset: IBM HR Analytics Employee Attrition & Performance

In the process of building a model, it's crucial to follow key steps:

1. Data Collection: Gather all necessary data.
2. Data Preparation: Clean, organize, and standardize data for accuracy.
3. Exploratory Data Analysis: Identify patterns and trends in the data.
4. Insight Generation: Extract actionable insights from analysis.
5. Decision Making: Utilize data-driven strategies to enhance workforce performance and meet organizational goals.

Steps Achieved:

1. Loaded the data using pandas library.
2. Performed data cleaning by checking if any null values, found the data to be clean.
3. Stage to do exploratory data analysis:
    - We dropped the unnecessary features from the data set.
    - Compared attrition with all categorical as well as continuous variable, insights generated as:
        1. Suggests a higher attrition rate among males compared to females.
        2. The age group between 28-32 witnesses the highest attrition rate, notably between 18-20, often see heightened attrition pattern then reaches a turning point around the age of 21.
        3. Significant increase in attrition rates specifically below 5000 per month. This trend gradually decreases, with a slight increase observed around the 10000 mark
        4. Employee with higher salary remain with company while other leave.
        5. Employee who started their career with company tend to leave job more than who share a good experience with the company.
4. Data preprocessing:

    1. Encoded the categorical variables to the numerical form using Labelencoding.

    2. There are high correlation between some features:

    MonthlyIncome - JobLevel
    YearsInCurrentRole - YearAtCompany
    YearWithCurrManager - YearsInCurrentRole
    TotalWorkingYears - JobLevel
    TotalWorkingYears - MonthlyIncome
    PercentSalaryHike – PerformanceRating

    3. Calculating imbalance ratio:

Imbalance Ratio (Retained Employees : Attrited Employees): 5.20 : 1
Data is biased for retained employees.
Using SMOTE to balance the data.

5. Splitting data to test and train
6. Calculating ANNOVA and chi-square value.
7. Plotted the AUC-ROC curve for the classification model.
8. Model selection:
   XGBoost, SVM, Logistice Regression, KNN, Decision Tree

9. Model evaluation:

	Model	Cross Validation Score	ROC-AUC	F1 Score (Attrition)	F1 Score (No Attrition)
0	XGBoost (XGB)	91.40%	84.10%	0.84	0.84
1	SVM	90.77%	83.53%	0.83	0.84
2	Decision Tree	91.74%	77.66%	0.77	0.78
3	Random Forest	88.19%	80.64%	0.80	0.82
4	KNN	89.19%	80.07%	0.81	0.78
5	Logistic Regression	87.50%	78.36%	0.78	0.79

Here we are getting most accuracy with XGBoost algorithm.

10. Summary:

- Gender disparity: Higher attrition rate among males compared to females
- Age dynamics: Highest attrition between ages 28-32, declining with age
- Income levels: Spikes in attrition at very low income, decreasing as income rises
- Job satisfaction: Lower satisfaction correlates with higher attrition, especially for average salaries
- Departmental differences: Sales and HR have highest attrition, R&D lower
- Job role impact: Higher-level roles have lower attrition rates
- Salary increment influence: Enhanced increments incentivize retention
- Educational background: Lower attrition among higher education levels
- Salary and stock options: Higher pay and stock options promote loyalty
- Work-life balance: Crucial factor affecting motivation and retention