Problem Statement

The objective is to predict employee attrition based on various employee attributes using machine learning models. The goal is to identify key factors contributing to attrition and develop a predictive model to help the company proactively manage employee retention.

Steps performed

Data Preprocessing:

I. First, we cleaned the dataset by removing unnecessary columns like 'EmployeeCount', 'StandardHours', 'EmployeeNumber', 'Over18', 'Unnamed: 35', 'Unnamed: 36'.

II. We handled categorical data in two ways: binary categories (like Gender, Attrition) were converted using Label Encoder, while multi-category variables (like Department, JobRole) were converted using One-Hot Encoding

III. Numerical features were standardized to have zero mean and unit variance using StandardScaler.

IV. Due to imbalanced attrition data (84% No vs 16% Yes), we used SMOTE to create synthetic samples of the minority class, achieving a 70:30 ratio.

Feature Engineering

I. Created binary (0/1) columns for each category in categorical variables through One-Hot Encoding

II. Scaled numerical features to ensure all variables contribute equally to the model

III. Data was split into 80% training and 20% testing sets while maintaining the same proportion of classes in both sets

Model Deployment using Streamlit

I. Created an interactive web application using Streamlit framework

II. Implemented user input forms for all required features (demographics, job-related factors, satisfaction levels, etc.)

III. Integrated the best performing model (CatBoost) into the application

IV. Added real-time prediction functionality where users can input employee details and get immediate attrition prediction

V. Included data validation and user-friendly interface for easy interaction

Key Findings from EDA

1. Department Distribution

- Research & Development: 65.4%
- Sales: 30.3%
- Human Resources: 4.3%

2. Job Satisfaction

- High satisfaction (level 4): 31.2%
- Medium satisfaction (level 3): 30.1%
- Lower satisfaction (levels 1-2): 38.7%

3. Work-Life Balance

- Good balance (level 3): 60.7%
- Better balance (level 4): 10.4%
- Poor balance (levels 1-2): 28.8%

4. Gender Distribution

- Male: 60%
- Female: 40%

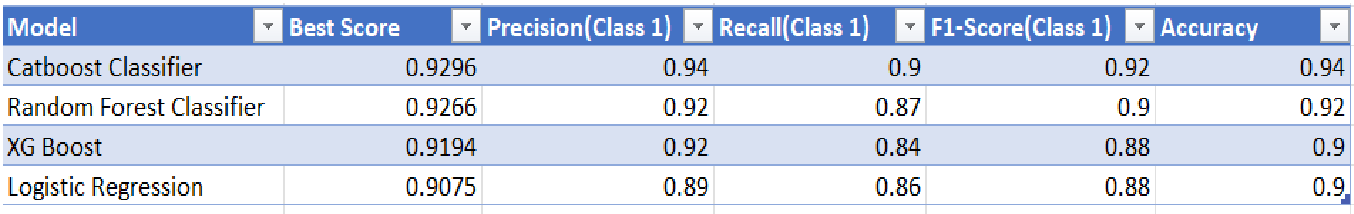
5. Attrition Distribution

- Significant class imbalance
- No attrition: ~84%
- Yes attrition: ~16%

6. Heatmap

- **TotalWorkingYears** - Shows a negative correlation with Attrition. Employees with more working years are less likely to leave.
- **YearsAtCompany** - Also negatively correlated, meaning employees who have spent more years at the company tend to stay.
- **YearsInCurrentRole** - Another negative correlation, indicating that employees in the same role for longer durations are less likely to leave.
- **YearsWithCurrManager** - Similar to the above, employees who have been with their current manager for longer durations are less likely to leave.
- **JobLevel & MonthlyIncome** - These factors also have a negative correlation, meaning employees with higher job levels and salaries are less likely to leave.

Model Comparison



Model	Best Score	Precision(Class 1)	Recall(Class 1)	F1-Score(Class 1)	Accuracy
Catboost Classifier	0.9296	0.94	0.9	0.92	0.94
Random Forest Classifier	0.9266	0.92	0.87	0.9	0.92
XG Boost	0.9194	0.92	0.84	0.88	0.9
Logistic Regression	0.9075	0.89	0.86	0.88	0.9

Final Conclusions and Recommendations

1. Model Selection

I. CatBoost performs best with 94% accuracy

II. Tree-based models consistently outperform Logistic Regression

2. Business Recommendations

I. Focus on R&D department which has highest employee concentration

II. Implement better work-life balance programs as 28.8% report poor balance

III. Address job satisfaction as 38.7% report lower satisfaction levels

IV. Consider regular employee surveys to monitor satisfaction levels

3. Future Improvement

I. Collect more data so better prediction can be made.

II. For different department consider different strategies.

III. Train model periodically to check for any problem and access the problem early as possible.