# MGMT 59000- Big Data and ML Ops (Final Project)

## *"IBM HR Analytics Employee Attrition"*

## Problem Statement: Attrition in an Organization – Why Workers Quit?

Employee attrition is a critical challenge for organizations, impacting productivity, profitability, and long-term growth. As employees form the backbone of any organization, their departure leads to significant disruptions, including:

- **High replacement costs** in terms of recruitment, training, and onboarding.
- **Loss of experienced employees**, reducing knowledge retention and skill depth.
- **Decreased productivity**, affecting overall team efficiency and project continuity.
- **Negative impact on profitability** due to lower employee engagement and increased hiring expenses.

This project aims to analyze the **IBM HR Analytics Attrition Dataset** to identify key factors contributing to employee turnover. By leveraging employee demographics, job-related data, compensation, work-life balance, and performance metrics, we seek to uncover trends and patterns in attrition.

The findings will help organizations **develop targeted retention strategies**, optimize HR policies, and enhance employee satisfaction, ultimately reducing workforce turnover.

## Dataset: "IBM HR Analytics Employee Attrition & Performance"

This is a fictional data set created by IBM data scientists. The **IBM HR Analytics Attrition Dataset** is a structured dataset aimed at understanding employee attrition within an organization. It contains detailed demographic, job-related, and performance metrics for employees, helping organizations analyze factors influencing turnover rates.

**Dataset Source:**

- **Provider**: IBM
- **Source**: Kaggle (IBM HR Analytics Dataset)
- **Use Case**: Employee attrition analysis for HR analytics

**Dataset Structure & Features:**

The dataset consists of **35 attributes**, categorized into **personal details, job-related information, work-life balance factors, compensation, and performance metrics**.

**Key Features**

1. **Demographics**
    - Age: Employee's age
    - Gender: Male/Female
    - MaritalStatus: Single/Married/Divorced
2. **Job & Employment Details**
    - JobRole: Specific job title (e.g., Sales Executive, R&D Scientist)
    - Department: Business unit (e.g., Sales, R&D, HR)
    - JobLevel: Seniority level in the company (1 to 5)
    - YearsAtCompany: Employee tenure
3. **Work-Life Balance & Compensation**
    - MonthlyIncome: Salary earned per month
    - OverTime: Whether the employee works overtime (Yes/No)
    - BusinessTravel: Travel frequency (Rarely, Frequently, None)
    - WorkLifeBalance: Rating (1 - Poor, 4 - Excellent)
4. **Performance & Career Growth**
    - PerformanceRating: Employee's performance evaluation (1 to 4)
    - TrainingTimesLastYear: Number of training programs attended
    - YearsSinceLastPromotion: Time since last career advancement
5. **Attrition Indicator**
    - Attrition: **Target variable (Yes/No)** indicating whether an employee left the company.

**Objective of Analysis**

This dataset is leveraged to analyze **attrition trends**, identifying key factors contributing to employee turnover. Insights gained can help organizations:

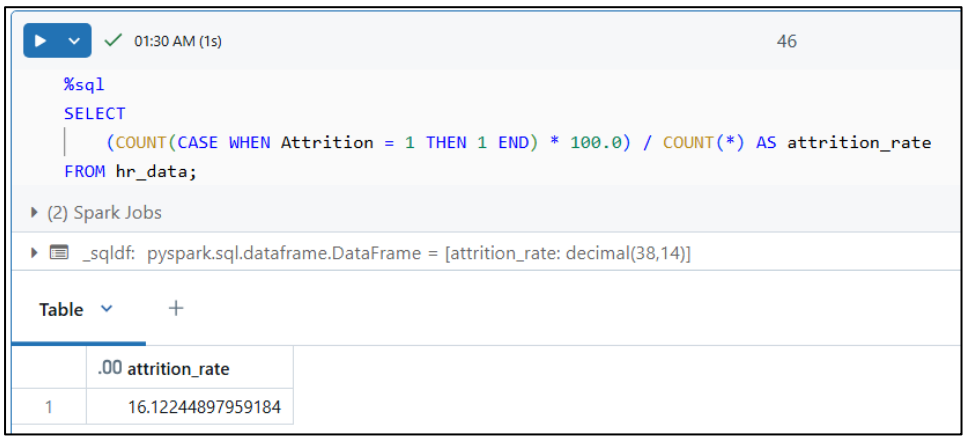
- Improve **employee retention strategies**
- Optimize **work-life balance and compensation policies**
- Identify **departments and job roles** with the highest attrition risks
- Develop **data-driven HR policies** to enhance employee satisfaction and reduce turnover

This dataset serves as a **real-world HR analytics case study**, providing actionable insights into employee behavior and organizational efficiency.

## SPARK SQL QUERIES (EDA) :

*These queries help us get preliminary insights from the dataset.*

**1. What is the attrition rate in the organization?**




An attrition rate of **16.12%** suggests a **moderate level of turnover**. While not alarmingly high, it's something the organization should monitor, especially if it is trending upwards over time.

## 2. Which department has the highest attrition rate?


- **Sales Department has the highest attrition rate (20.63%)**, likely due to high-pressure targets, commission-based pay, and burnout.
- **HR follows closely (19.05%)**, which may indicate dissatisfaction with policies, lack of career growth, or workload issues.
- **R&D has lower attrition (13.84%)**, suggesting better job satisfaction, stability, or specialized skill retention.
- **Action Needed:** Focus on retention strategies for Sales & HR through better incentives, work-life balance, and engagement initiatives.

## 3. How does job level impact attrition?


- **Lower job levels face higher attrition** – Job Level **1 has the highest attrition (26.34%)**, indicating dissatisfaction with pay, growth, or job stability.
- **Mid-Level employees (Level 3) see a moderate rate (14.67%)**, suggesting some career mobility but still room for improvement.

- **Higher job levels (4 & 5) have the lowest attrition (~7%)**, likely due to better compensation, stability, and job satisfaction.
- **Action Needed** – Focus on **career progression, mentorship, and pay raises for entry-level employees** to reduce early-stage turnover.

### 4. What is the impact of business travel on attrition?


- **Frequent travelers have the highest attrition (24.91%)**, likely due to work-life balance challenges, stress, and exhaustion.
- **Employees who travel rarely have a moderate attrition rate (14.96%)**, suggesting some impact but not as severe.
- **Non-travelers have the lowest attrition (8%)**, indicating more stability and job satisfaction.
- **Action Needed** – Implement better **travel policies, flexible schedules, and additional benefits** to retain frequent travelers.

### 5. Does marital status affect attrition?

- **Singles have the highest attrition rate (25.53%)**, likely due to greater flexibility to switch jobs or relocate for better opportunities.
- **Married employees have a lower attrition rate (12.48%)**, suggesting job stability due to family commitments and financial responsibilities.
- **Divorced employees show the lowest attrition (10.09%)**, possibly valuing job security and benefits more.
- **Action Needed** – Offer **incentives, career growth opportunities, and retention programs for single employees** to reduce turnover.

### 6. How does work-life balance affect attrition?

```
%sql
SELECT WorkLifeBalance,
       COUNT(CASE WHEN Attrition = 1 THEN 1 END) * 100.0 / COUNT(*) AS attrition_rate
FROM hr_data
GROUP BY WorkLifeBalance
ORDER BY WorkLifeBalance;
```

01:42 AM (1s) 56

(2) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [WorkLifeBalance: integer, attrition_rate: decimal(38,14)]

Table

| | WorkLifeBalance | attrition_rate |
|---|---|---|
| 1 | 1 | 31.25000000000000 |
| 2 | 2 | 16.86046511627907 |
| 3 | 3 | 14.22172452407615 |
| 4 | 4 | 17.64705882352941 |

4 rows | 0.51s runtime Refreshed 17 hours ago

This result is stored as `_sqldf` and can be used in other Python cells.

- **Poor work-life balance (rating 1) has the highest attrition (31.25%)**, indicating employees struggle with long hours, stress, or lack of flexibility.
- **Moderate work-life balance (ratings 2 & 3) sees lower attrition (16.86% and 14.22%)**, suggesting some dissatisfaction but manageable.
- **Surprisingly, the best work-life balance (rating 4) still has 17.65% attrition**, which might indicate other job-related factors at play.
- **Action Needed** – Implement **flexible schedules, remote work options, and mental health support** to retain employees with poor work-life balance.

## 7. What is the average monthly income of employees who left vs. stayed?

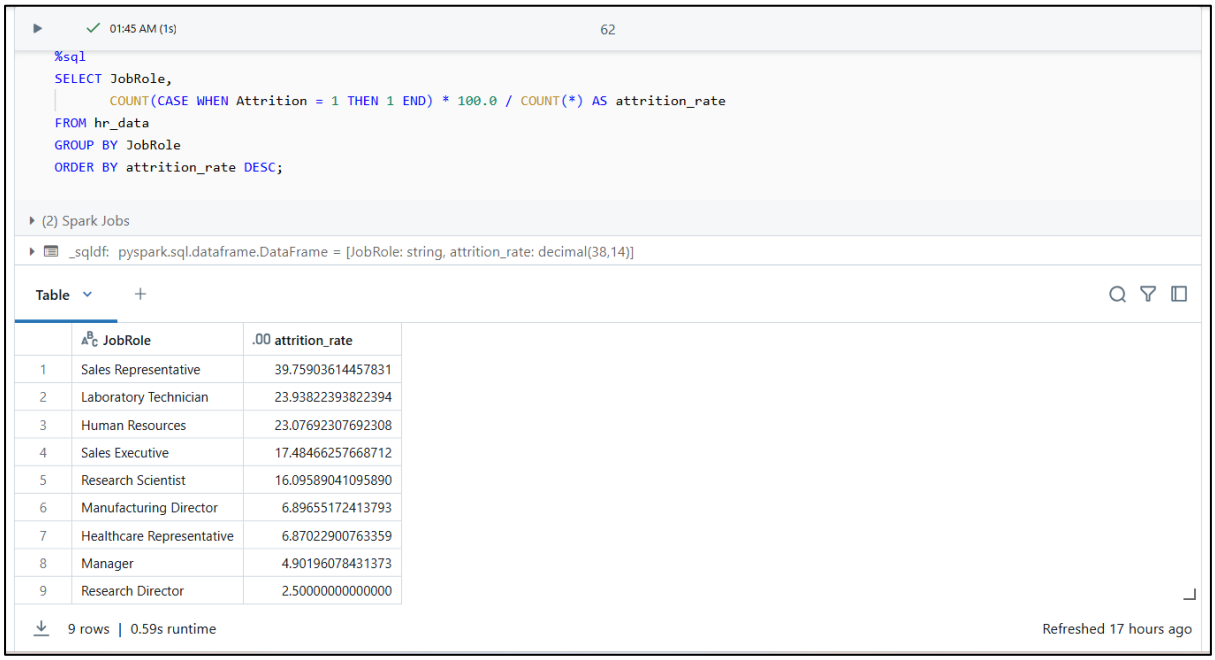

- **Employees who left had a significantly lower average income ($4,787)** compared to those who stayed ($6,832), indicating salary dissatisfaction as a key factor for attrition.
- **Low pay may be driving resignations**, especially for entry-level employees or departments with high turnover.
- **Action Needed** – Conduct **salary benchmarking, offer competitive raises, and implement performance-based incentives** to retain employees.

## 8. How does overtime impact attrition?


- **Employees working overtime have a much higher attrition rate (30.53%)**, indicating burnout, stress, and poor work-life balance.

- **Employees without overtime have a significantly lower attrition rate (10.44%)**, suggesting better job satisfaction and stability.
- **Action Needed** – Implement **workload management, fair overtime compensation, and mandatory rest periods** to reduce turnover among overworked employees.

### 9. What job roles have the highest attrition?



```
%sql
SELECT JobRole,
       COUNT(CASE WHEN Attrition = 1 THEN 1 END) * 100.0 / COUNT(*) AS attrition_rate
FROM hr_data
GROUP BY JobRole
ORDER BY attrition_rate DESC;
```

01:45 AM (1s) 62

(2) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [JobRole: string, attrition_rate: decimal(38,14)]

Table

| | JobRole | attrition_rate |
|---|---|---|
| 1 | Sales Representative | 39.75903614457831 |
| 2 | Laboratory Technician | 23.93822393822394 |
| 3 | Human Resources | 23.07692307692308 |
| 4 | Sales Executive | 17.48466257668712 |
| 5 | Research Scientist | 16.09589041095890 |
| 6 | Manufacturing Director | 6.89655172413793 |
| 7 | Healthcare Representative | 6.87022900763359 |
| 8 | Manager | 4.90196078431373 |
| 9 | Research Director | 2.50000000000000 |

9 rows | 0.59s runtime Refreshed 17 hours ago

- **Sales Representatives have the highest attrition (39.76%)**, likely due to high-pressure targets, commission-based pay, and job stress.
- **Laboratory Technicians (23.94%) and HR (23.08%)** also face significant attrition, possibly due to workload issues or limited career growth.
- **Sales Executives (17.48%) and Research Scientists (16.10%)** show moderate attrition, indicating some dissatisfaction.
- **Action Needed** – Focus on **better incentives, career growth plans, and work-life balance improvements** for high-attrition roles, especially in Sales and HR.

### 10. How does total working experience influence attrition?

```
%sql
SELECT TotalWorkingYears,
       COUNT(CASE WHEN Attrition = 1 THEN 1 END) * 100.0 / COUNT(*) AS attrition_rate
FROM hr_data
GROUP BY TotalWorkingYears
ORDER BY TotalWorkingYears;
```

(2) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [TotalWorkingYears: integer, attrition_rate: decimal(38,14)]

Table

| | TotalWorkingYears | attrition_rate |
|---|---|---|
| 1 | 0 | 45.45454545454545 |
| 2 | 1 | 49.38271604938272 |
| 3 | 2 | 29.03225806451613 |
| 4 | 3 | 21.42857142857143 |
| 5 | 4 | 19.04761904761905 |
| 6 | 5 | 18.18181818181818 |
| 7 | 6 | 17.60000000000000 |
| 8 | 7 | 22.22222222222222 |
| 9 | 8 | 15.53398058252427 |
| 10 | 9 | 10.41666666666667 |
| 11 | 10 | 12.37623762376238 |

- **Employees with 0-1 years of experience have the highest attrition (45-49%)**, likely due to job dissatisfaction, lack of career growth, or better external opportunities.
- **Attrition decreases as experience increases**, dropping to **18-19% for employees with 4-5 years of experience**, suggesting greater stability.
- **Action Needed** – Improve **onboarding, career development programs, and mentorship** to retain early-career employees and reduce high turnover in the first year.

## A few complex questions to gain some depth !!

### 1. How does a combination of Job Role, OverTime, and Job Satisfaction impact attrition?

```
-- This query helps analyze how overtime and job satisfaction affect employees in different job roles, giving a multi-dimensional insight.

SELECT JobRole, OverTime, JobSatisfaction,
       COUNT(CASE WHEN Attrition = 1 THEN 1 END) AS attrition_count,
       COUNT(*) AS total_employees,
       (COUNT(CASE WHEN Attrition = 1 THEN 1 END) * 100.0) / COUNT(*) AS attrition_rate
FROM hr_data
GROUP BY JobRole, OverTime, JobSatisfaction
ORDER BY attrition_rate DESC;
```

(2) Spark Jobs

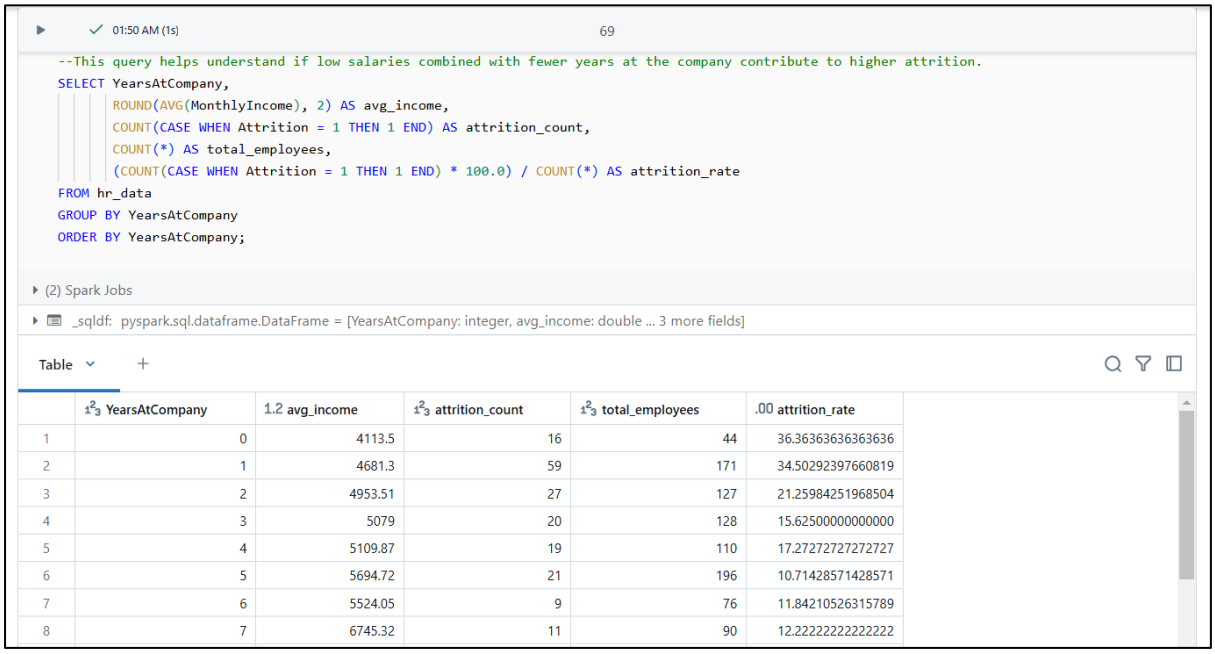
_sqldf: pyspark.sql.dataframe.DataFrame = [JobRole: string, OverTime: string ... 4 more fields]

Table

| | JobRole | OverTime | JobSatisfaction | attrition_count | total_employees | attrition_rate |
|---|---|---|---|---|---|---|
| 1 | Sales Representative | Yes | 2 | 7 | 7 | 100.00000000000000 |
| 2 | Human Resources | Yes | 1 | 3 | 3 | 100.00000000000000 |
| 3 | Laboratory Technician | Yes | 2 | 3 | 4 | 75.00000000000000 |
| 4 | Sales Representative | No | 1 | 7 | 11 | 63.63636363636364 |
| 5 | Sales Representative | Yes | 4 | 4 | 7 | 57.14285714285714 |
| 6 | Laboratory Technician | Yes | 3 | 10 | 18 | 55.55555555555556 |
| 7 | Sales Representative | Yes | 3 | 5 | 9 | 55.55555555555556 |
| 8 | Human Resources | Yes | 3 | 1 | 2 | 50.00000000000000 |

- **Sales Representatives and HR roles with overtime and low job satisfaction have the highest attrition (up to 100%)**, indicating extreme burnout and dissatisfaction.
- **Laboratory Technicians with overtime also experience high attrition (75-55%)**, suggesting workload stress in technical roles.
- **Sales Representatives with no overtime still face high attrition (63.64%)**, implying job-related stress beyond overtime factors.
- **Action Needed** – HR should **reduce excessive overtime, improve job satisfaction initiatives, and offer retention incentives** for high-risk roles like Sales, HR, and Technical positions.

**2. What is the correlation between Monthly Income, Years at Company, and Attrition?**



```
01:50 AM (1s)                                                    69
--This query helps understand if low salaries combined with fewer years at the company contribute to higher attrition.
SELECT YearsAtCompany,
       ROUND(AVG(MonthlyIncome), 2) AS avg_income,
       COUNT(CASE WHEN Attrition = 1 THEN 1 END) AS attrition_count,
       COUNT(*) AS total_employees,
       (COUNT(CASE WHEN Attrition = 1 THEN 1 END) * 100.0) / COUNT(*) AS attrition_rate
FROM hr_data
GROUP BY YearsAtCompany
ORDER BY YearsAtCompany;

(2) Spark Jobs
_sqldf: pyspark.sql.dataframe.DataFrame = [YearsAtCompany: integer, avg_income: double ... 3 more fields]
```

Table

| | YearsAtCompany | avg_income | attrition_count | total_employees | attrition_rate |
|---|---|---|---|---|---|
| 1 | 0 | 4113.5 | 16 | 44 | 36.363636363636363636 |
| 2 | 1 | 4681.3 | 59 | 171 | 34.50292397660819 |
| 3 | 2 | 4953.51 | 27 | 127 | 21.25984251968504 |
| 4 | 3 | 5079 | 20 | 128 | 15.62500000000000 |
| 5 | 4 | 5109.87 | 19 | 110 | 17.27272727272727 |
| 6 | 5 | 5694.72 | 21 | 196 | 10.71428571428571 |
| 7 | 6 | 5524.05 | 9 | 76 | 11.84210526315789 |
| 8 | 7 | 6745.32 | 11 | 90 | 12.22222222222222 |

- **Newer employees (0-2 years) with lower salaries have the highest attrition (36-34%)**, indicating early job dissatisfaction or better external opportunities.
- **Attrition decreases as tenure increases**, dropping to **~11-12% for employees with 7+ years**, suggesting long-term employees find stability.
- **Salary gradually increases with tenure**, implying retention is linked to financial growth.
- **Action Needed** – Implement **early-stage salary adjustments, career development programs, and stronger onboarding support** to retain newer employees.

### 3. Which business unit is losing its high performers?

```sql
--This query identifies departments that are losing top-performing employees, which could indicate dissatisfaction among high-achieving
employees.

SELECT Department,
       PerformanceRating,
       COUNT(CASE WHEN Attrition = 1 THEN 1 END) AS high_performer_attrition_count,
       COUNT(*) AS total_high_performers,
       (COUNT(CASE WHEN Attrition = 1 THEN 1 END) * 100.0) / COUNT(*) AS high_performer_attrition_rate
FROM hr_data
WHERE PerformanceRating >= 4
GROUP BY Department, PerformanceRating
ORDER BY high_performer_attrition_rate DESC;
```

(2) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [Department: string, PerformanceRating: integer ... 3 more fields]

Table

| | Department | PerformanceRating | high_performer_attrition_count | total_high_performers | high_performer_attrition_rate |
|---|---|---|---|---|---|
| 1 | Research & Developme... | 4 | 26 | 156 | 16.66666666666667 |
| 2 | Sales | 4 | 10 | 61 | 16.39344262295082 |
| 3 | Human Resources | 4 | 1 | 9 | 11.11111111111111 |

- **Research & Development (16.67%) and Sales (16.39%) are losing the most high-performing employees**, indicating dissatisfaction with career growth, leadership, or incentives.
- **HR follows with 11.11% attrition of high performers**, which may signal issues with work culture or lack of internal career advancement.
- **Losing top talent can hurt innovation, revenue, and overall performance** if not addressed proactively.
- **Action Needed** – HR should **offer better career progression, retention bonuses, and leadership development programs** to retain high performers in R&D and Sales.

# Machine Learning: Predicting Employee Attrition Using Spark MLlib

### 1. Objective

The goal of this study is to develop **machine learning models** to predict employee attrition using the **IBM HR Analytics Dataset**. By identifying key factors influencing employee turnover, businesses can implement proactive HR strategies to **retain talent** and reduce attrition-related costs.

### 2. Data Preprocessing for Machine Learning

To ensure high-quality input for machine learning models, the following preprocessing steps were performed:

- **Categorical Encoding:**

- Used **StringIndexer** and **OneHotEncoder** to convert categorical variables into numerical format.

- **Feature Engineering:**
  - Employed **VectorAssembler** to combine all relevant features into a **single feature vector**.
- **Train-Test Split:**
  - The dataset was split into **80% training** and **20% testing** to evaluate model performance effectively.

**3. Models Implemented**

Four classification models were implemented using **Spark MLlib** to predict attrition:

1. **Logistic Regression:**
   - Baseline model used for benchmarking.
2. **Random Forest Classifier:**
   - Strong tree-based model, effective for structured data.
3. **XGBoost:**
   - Advanced boosting algorithm, widely used for handling **tabular data**.
4. **CatBoost:**
   - Best suited for datasets with **heavy categorical features**, which makes it an ideal choice for HR data.

- **Pipeline Approach:**
  - Feature transformations and model training were automated using **ML Pipelines**.

**4. Feature Importance**

The most significant factors affecting employee attrition were identified:

- **Overtime:** Employees **without overtime** were more likely to leave.
- **Monthly Income:** Lower income employees exhibited higher attrition rates.
- **Age:** Younger employees had higher turnover rates.

These insights provide actionable steps for HR teams to **optimize compensation, work-life balance, and retention strategies**.

**5. Model Performance Evaluation**

To determine the best model for predicting attrition, the following metrics were analyzed:

- **Accuracy:** Measures overall correctness.
- **F1-Score:** Balances precision and recall (useful for imbalanced datasets).
- **AUC (Area Under ROC Curve):** Assesses the model's ability to differentiate between employees who stay and leave.
- **Precision-Recall Score:** Evaluates model effectiveness for high-risk attrition case.

**6. Model Comparison & Best Performing Model**

| Model | Accuracy | AUC (ROC) | Precision-Recall AUC | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 85.81% | 0.79 | 0.58 | **0.8581** |
| Random Forest | 87.32% | 0.81 | 0.61 | 0.8457 |
| XGBoost | 87.95% | 0.81 | 0.62 | 0.8523 |
| **CatBoost (Best Model)** | **88.98%** | **0.8227** | **0.6330** | 0.8500 |

- **CatBoost performed best overall** with an **88.98% accuracy** and **highest AUC (0.8227)**.
- It was also the most effective at detecting **high-risk employees likely to leave**.
- **Logistic Regression had the highest F1-score** but lacked interpretability compared to tree-based models.

**7. Final Takeaways & Business Implications**

- **Younger employees** (early-career professionals) are more likely to switch jobs, emphasizing the need for **career growth incentives**.
- **Salary and stock options** significantly impact retention—higher compensation correlates with employee loyalty.
- **Work-life balance** matters, but some employees still switch for better opportunities despite good conditions.
- **Sales and target-driven roles** show higher attrition, requiring better **incentive structures** and **stress management policies**.

By leveraging **machine learning insights**, HR teams can take **proactive measures** to reduce employee turnover and enhance workforce stability.

# MLOps Best Practices (ML Flow):

MLflow Steps for Model Tracking in Databricks:

**1. Setting Up MLflow Experiment**

- Defined an MLflow experiment: Users/your_email@databricks.com/Employee_Attrition_Prediction.
- Ensured the correct experiment path in Databricks.

**2. Logging Model Training Runs**

- Trained multiple models (Logistic Regression, Random Forest).
- Logged **model parameters, accuracy, and AUC**.
- Stored trained models for future evaluation.

**3. Tracking Runs in Databricks MLflow UI**

- The MLflow dashboard displays **each run's status, duration, and metrics.**
- ✅ **Green:** Successful model run.
- ❌ **Red:** Failed run due to missing setup or errors.

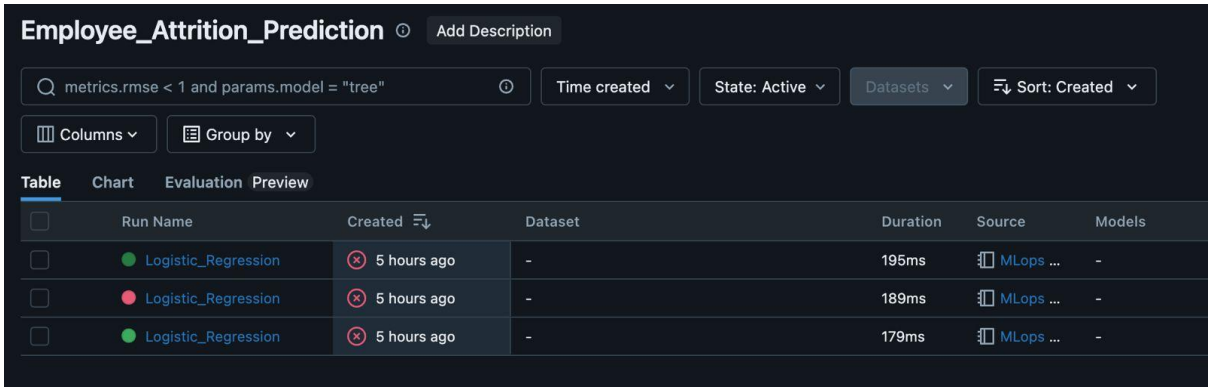
**4. Querying & Comparing Model Runs**

- Used filters like metrics.rmse < 1 and params.model = "tree" to find the best models.
- Allowed **comparison of different experiments** for performance evaluation.

**5. Next Steps**

- Extend MLflow tracking to **XGBoost & CatBoost**.
- Optimize models with **hyperparameter tuning** and log different configurations.
- Deploy the best-performing model in Databricks.

**MLflow ensures reproducibility, model tracking, and performance comparison in a scalable way.**



# Recommendation and Inferences:

Employee attrition is a dynamic challenge that affects organizations across industries. Based on our **machine learning analysis using Spark MLlib**, we derived key insights into the driving factors behind employee turnover.

**1. Age & Career Stability**

- Employees are more likely to **switch jobs early in their careers** or within the first few years of employment.
- Once employees establish **stability (marriage, family, career growth)**, they tend to stay longer within the same organization.

**2. Financial Incentives Matter**

- **Higher salaries and stock options** play a significant role in employee retention.
- Organizations that provide **competitive compensation** experience lower attrition rates.

**3. Work-Life Balance & Job Switching**

- Employees prioritize **work-life balance**, but even those with good balance may leave for **better opportunities** or career advancements.
- Ensuring **flexible schedules, mental health support, and career progression** can help mitigate this risk.

**4. Work Stress & Departmental Differences**

- Departments that have **high-pressure performance targets** (e.g., **Sales**) tend to experience **higher turnover**.
- Administrative and support roles (e.g., **HR, R&D**) face comparatively lower attrition rates.

**Final Takeaways:**

- **Proactive HR strategies**, such as **salary benchmarking, targeted retention programs, and employee engagement initiatives**, are critical in reducing attrition.
- **Data-driven insights from predictive modeling** can help businesses **identify high-risk employees** and take timely actions to retain top talent.

This study highlights the **importance of a structured approach** in addressing attrition, ensuring long-term organizational growth and workforce stability.