HR Attrition Risk Analysis & Workforce Segmentation

PROJECT REPORT

BY AYUSHMAN CHAKRABORTY

📝 Problem Statement

The company's leadership team was facing a significant business challenge with high employee turnover. While the overall headcount was stable, the consistent loss of talent was costly and disruptive. The HR department lacked a data-driven approach to understand the root causes and, more importantly, had no way to proactively identify which employees were at the highest risk of leaving.

Key business questions addressed:

- What are the key drivers of overall employee attrition?
- Which departments, job roles, and demographics are most affected?
- Can we build a reliable model to predict which employees are likely to leave in the future?
- Can we segment the workforce into distinct personas to better understand their needs and risk profiles?

The goal of this project was to build an end-to-end data science pipeline to answer these questions, culminating in an interactive dashboard featuring a predictive tool to help HR move from a reactive to a proactive retention strategy.

🛠️ Tech Stack

- **Data Storage & Querying:** SQLite
- **Data Analysis & Modeling:** Python
    - **Libraries:** Pandas (Data Manipulation), Scikit-learn (Classification & Clustering), Imblearn (SMOTE for imbalance), Matplotlib & Seaborn (Visualization)
- **Business Intelligence & Visualization:** Power BI
- **Dataset:** IBM HR Analytics Employee Attrition & Performance (Kaggle)

📂 Project Workflow

1. **Data Cleaning & Feature Engineering (Python & Pandas)**
    - Loaded and profiled the raw dataset to identify and handle inconsistencies.
    - Performed cleaning tasks: standardized column names and dropped constant/irrelevant columns.
    - Engineered new, insightful features like tenure_bucket, salary_band, and composite_satisfaction to enrich the analysis.
    - Produced cleaned_hr_data.csv as the single source of truth for all subsequent analysis.

2. **Database Querying & EDA (SQL)**
    - Imported the cleaned data into an SQLite database for structured querying.
    - Conducted Exploratory Data Analysis (EDA) to calculate baseline attrition rates across key business segments like Department, Job Role, and Salary Band.
    - Used SQL to perform cohort analysis and demographic breakdowns.

3. **Machine Learning Modeling & Optimization (Python)**
    - **Employee Segmentation:** Applied K-Means Clustering on key employee attributes (satisfaction, income, tenure) to identify 4 distinct workforce personas.
    - **Attrition Prediction:** Developed a comprehensive workflow to train a Random Forest Classifier to predict employee attrition.
        - Handled class imbalance with SMOTE to ensure the model could effectively learn from the minority class (employees who left).
        - Implemented an advanced tuning pipeline including Feature Selection and Decision Threshold Adjustment to maximize the model's recall.

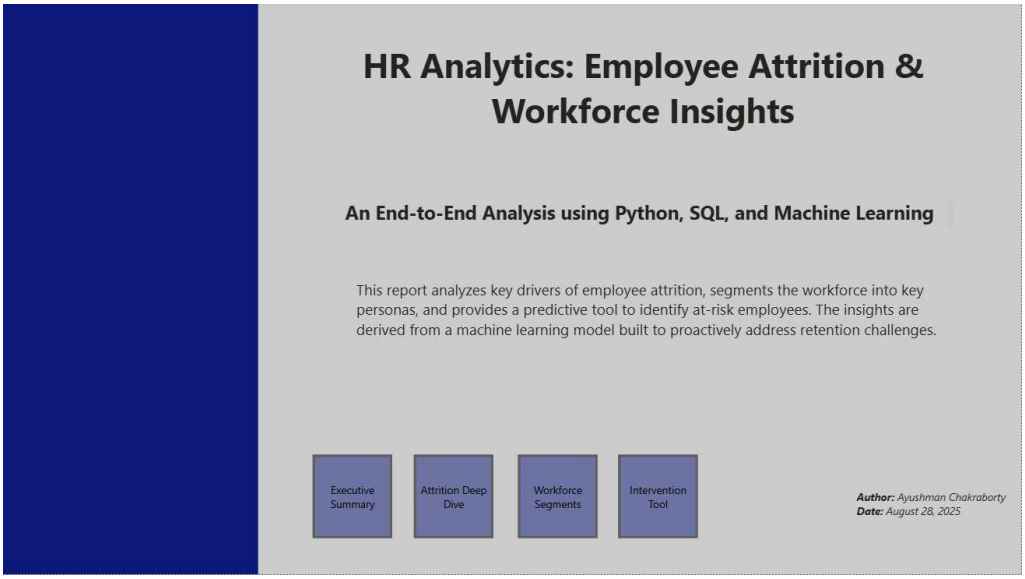
4. **Interactive Visualization (Power BI)**
    - Integrated the cleaned data and all machine learning outputs (clusters, predictions, risk scores) into Power BI.
    - Designed a 5-page interactive dashboard to tell a complete story, from a high-level overview to a predictive, actionable tool.

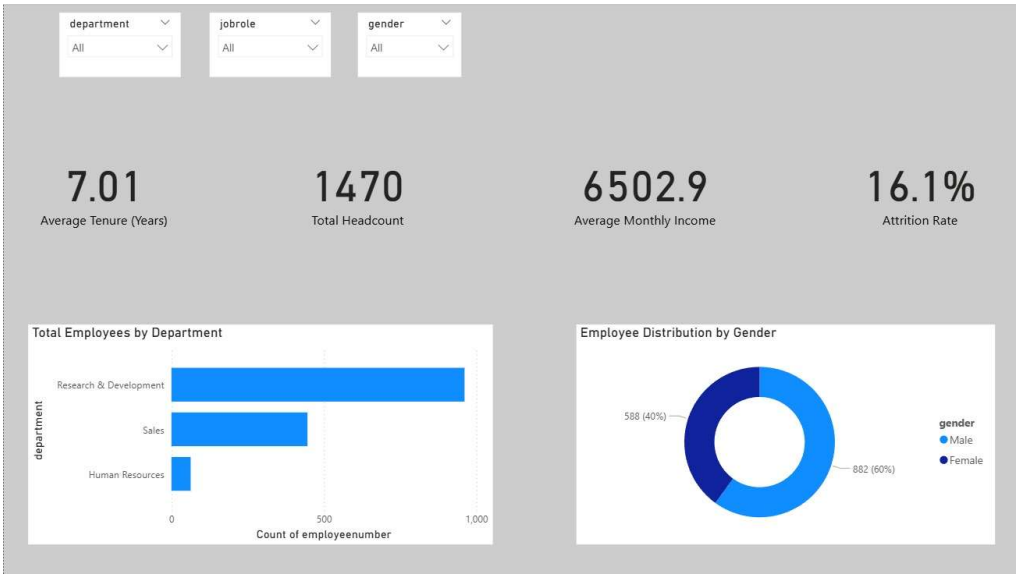
💡 **Dashboard Solution & Implementation**

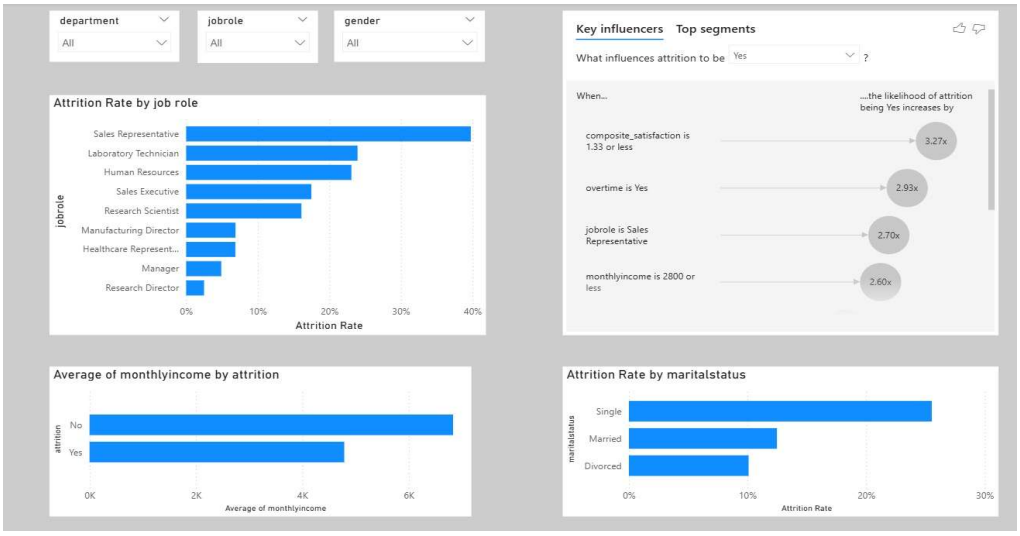
- **Page 1: Title Page**
    - **Purpose:** A professional introduction to the report, its objectives, and navigation.
    - **Implementation:** Clean layout with clear titles, a project summary, and a navigation bar with buttons for each report page.
- **Page 2: Executive Summary**
    - **Purpose:** Provide leadership with at-a-glance KPIs and a high-level overview of the workforce.
    - **Implementation:**
        - **KPI Cards:** Total Headcount, Overall Attrition Rate, Average Tenure, Average Monthly Income.
        - **Charts:** Bar chart for headcount by department and a doughnut chart for gender distribution.
        - **Interactivity:** Dropdown slicers for Department, Job Role, and Gender to filter the entire page.
- **Page 3: Attrition Deep Dive**
    - **Purpose:** A diagnostic tool to explore the root causes of historical attrition.
    - **Implementation:**
        - **Bar Charts:** Attrition Rate by Job Role and Marital Status, sorted to highlight problem areas.
        - **AI Visual:** A Key Influencers chart to automatically identify the top factors driving attrition.
        - **Comparative Analysis:** A bar chart comparing the average monthly income of employees who left versus those who stayed.
- **Page 4: Workforce Segmentation**
    - **Purpose:** A detailed visualization of the 4 employee personas discovered by the K-Means model.
    - **Implementation:**
        - **Scatter Plot:** Visualizing the clusters based on income vs. satisfaction.
        - **Bar Charts:** Showing the size (headcount) and risk level (attrition rate) of each segment.
        - **Summary Table:** A detailed table of the average characteristics for each named persona.

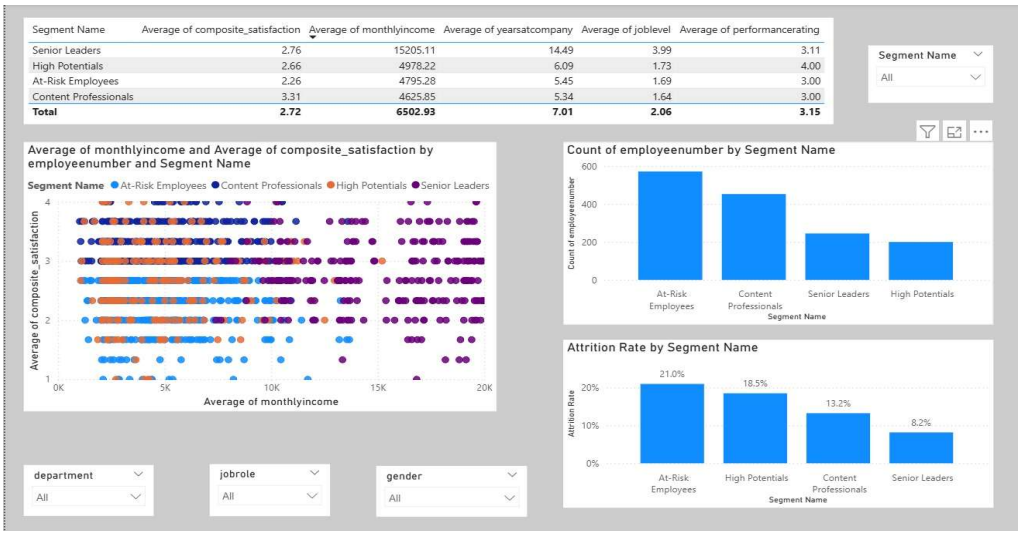
- **Page 5: Retention Intervention Tool**
    - **Purpose:** The final, predictive page, providing a prioritized list of employees flagged by the model as high flight risks.
    - **Implementation:**
        - **Page-Level Filter:** The entire page is filtered to show only employees with a predicted_attrition of 1.
        - **Prioritized List:** A detailed table of high-risk employees, sorted descending by their attrition_risk_score.
        - **Summary Visuals:** KPI cards and charts summarizing the high-risk group.

DASHBOARD SNAPSHOTS



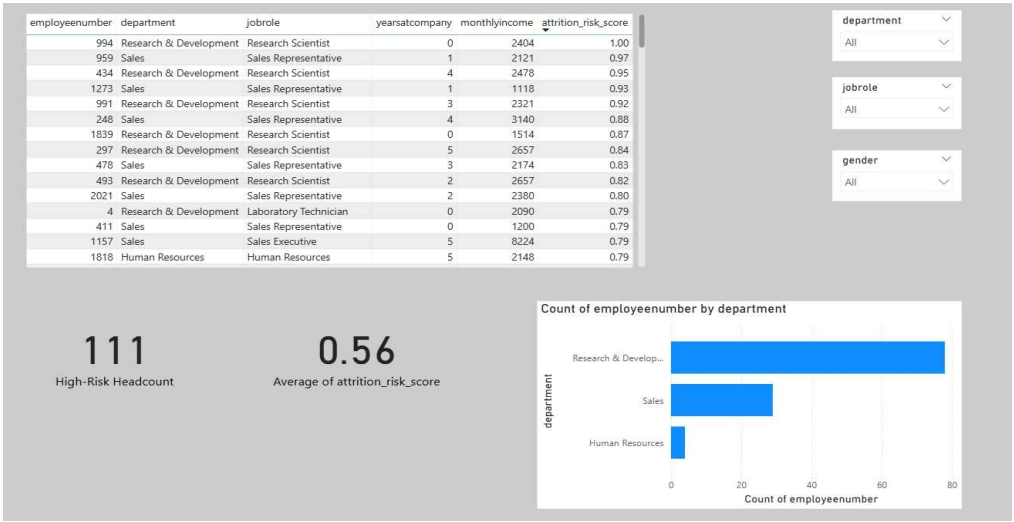






Segment Name	Average of composite_satisfaction	Average of monthlyincome	Average of yearsatcompany	Average of joblevel	Average of performancerating
Senior Leaders	2.76	15205.11	14.49	3.99	3.11
High Potentials	2.66	4978.22	6.09	1.73	4.00
At-Risk Employees	2.26	4795.28	5.45	1.69	3.00
Content Professionals	3.31	4625.85	5.34	1.64	3.00
Total	**2.72**	**6502.93**	**7.01**	**2.06**	**3.15**

employeenumber	department	jobrole	yearsatcompany	monthlyincome	attrition_risk_score
994	Research & Development	Research Scientist	0	2404	1.00
959	Sales	Sales Representative	1	2121	0.97
434	Research & Development	Research Scientist	4	2478	0.95
1273	Sales	Sales Representative	1	1118	0.93
991	Research & Development	Research Scientist	3	2321	0.92
248	Sales	Sales Representative	4	3140	0.88
1839	Research & Development	Research Scientist	0	1514	0.87
297	Research & Development	Research Scientist	5	2657	0.84
478	Sales	Sales Representative	3	2174	0.83
493	Research & Development	Research Scientist	2	2657	0.82
2021	Sales	Sales Representative	2	2380	0.80
4	Research & Development	Laboratory Technician	0	2090	0.79
411	Sales	Sales Representative	0	1200	0.79
1157	Sales	Sales Executive	5	8224	0.79
1818	Human Resources	Human Resources	5	2148	0.79



Detailed Analysis & Findings

1. Foundational Analysis: Who is Leaving? (SQL & EDA)

The initial analysis focused on understanding the high-level patterns of historical attrition to identify the most affected areas of the business.

- **Overall Attrition:** The company has a baseline attrition rate of **16.1%**. This figure serves as a critical benchmark; any group with an attrition rate significantly higher than this is a key area of concern.
- **Job Role Impact:** Attrition is far from uniform across the company. The data clearly shows that front-line, entry-level roles are the most volatile. **Sales Representatives** exhibit the highest turnover rate by a wide margin (over 39%), followed by **Laboratory Technicians** and **Human Resources** staff. Conversely, senior and managerial roles show high stability, with very low attrition rates. This suggests that career progression and role seniority are strong factors in employee retention.
- **Demographic Impact:** Among demographic factors, marital status was the most significant signal. **Single** employees are considerably more likely to leave than their Married or Divorced colleagues. This insight can help HR tailor retention programs, potentially focusing on building a stronger sense of community and stability for younger, unattached employees.

2. Predictive Modeling: Why Are They Leaving?

To move from "who" to "why," a Random Forest classification model was built and rigorously optimized. The final model's primary goal was to act as an effective early-warning system. After a full tuning pipeline—including feature selection and threshold adjustment—the model was successfully optimized to have an accuracy of 78 percent and correctly identify 61% (Recall) of all employees who would actually leave, making it a powerful tool for proactive intervention.

The feature importance analysis from the model provided a clear hierarchy of attrition drivers:

1. **Overtime:** This was the single most powerful predictor. Employees who work overtime are dramatically more likely to leave, suggesting that burnout and work-life imbalance are critical issues that need to be addressed.
2. **Monthly Income:** As expected, compensation is a major factor. The model confirmed that lower income is a strong predictor of turnover, especially for employees in junior roles.
3. **Job Level & Age:** These two factors are closely related. The model found that employees in more junior roles (lower job levels) and those who are younger are at a significantly higher risk of leaving, highlighting a retention challenge with early-career talent.
4. **Composite Satisfaction:** The satisfaction score we engineered proved to be a valuable predictor. As expected, employees with lower overall satisfaction across their job, environment, and relationships were more likely to churn.

3. Workforce Intelligence: The Four Employee Personas

The K-Means clustering model successfully segmented the workforce into four distinct, actionable personas. Analyzing these groups provides a strategic, high-level view of the workforce's structure and its inherent risks and strengths.

- **Segment 1: Senior Leaders (16% of workforce):**
    - **Characteristics:** This segment is defined by seniority and experience. They have, by far, the highest average monthly income (~$15,200), the longest tenure (~14.5 years), and the highest job levels. Their satisfaction is reasonably high.
    - **Business Value:** This is the experienced leadership and expert core of the company. Their institutional knowledge and stability are invaluable. Their attrition rate is the **lowest in the company (8.2%)**, which is a positive sign, but retaining this group must remain a top priority due to the high cost of replacing senior talent.
- **Segment 2: At-Risk Employees (38% of workforce):**
    - **Characteristics:** This is the largest segment, characterized by low income (~$4,800), junior roles, and, most critically, the **absolute lowest average satisfaction scores (2.26)**.
    - **Business Value:** This group represents the company's biggest vulnerability. Their combination of low satisfaction and junior status results in the **highest attrition rate (21.0%)**. They are actively disengaged and represent a significant, ongoing cost in recruitment and training. The "Intervention Tool" is primarily focused on identifying the highest-risk individuals within this critical segment before they resign.
- **Segment 3: High Potentials (13% of workforce):**
    - **Characteristics:** This segment consists of employees in junior, lower-paying roles, but they are distinguished by having the **highest average performance ratings (4.00)**. Despite their strong performance, their satisfaction levels are worryingly low (2.66).
    - **Business Value:** These are the company's rising stars and future leaders. Their high attrition rate **(18.5%)** is a major strategic concern, as it signals that the company is failing to retain its top-performing junior talent. This group likely feels undervalued and may be leaving for better growth opportunities and compensation elsewhere.
- **Segment 4: Content Professionals (31% of workforce):**
    - **Characteristics:** This group is the backbone of the company's daily operations. They are in junior, lower-paying roles but are distinguished by having the **highest satisfaction scores (3.31)**.
    - **Business Value:** This is the stable, happy core of the workforce. Their attrition rate is relatively low **(13.2%)**. The strategic goal for this segment is to maintain their high satisfaction and provide clear career paths to prevent them from becoming stagnant and eventually migrating into the "At-Risk" category.

This detailed analysis, from foundational EDA to predictive modeling and segmentation, provides the company with a comprehensive, data-driven understanding of its workforce and a clear path forward to improving employee retention.