

# Project 8 : SQL + ML Integration

## Objective:

Combine **SQL querying power** with **Machine Learning models** to analyze attrition risk. This project demonstrates how HR teams can query their employee database directly and run predictions on-the-fly, bridging People Analytics with HRIS-like systems.

## Why It Matters:

- HR data often lives in databases (HRIS, payroll systems).
- Analysts should be able to run queries and pipe results into ML models.
- This integration makes predictive attrition analytics more practical in enterprise contexts.

✓ DB already exists: C:\Users\amlanmishra2\hr\_dataset.db  
Tables: []

✓ Created hr\_dataset.db with table 'employees' (1470 rows). Location: C:\Users\amlanmishra2\hr\_dataset.db

## Attrition by Department

Out[33]:

	Department	total	left_count
0	Research & Development	9 6 1	1 3 3
1	Sales	4 4 6	9 2
2	Human Resources	6 3	1 2

## Sample Query (sanity check)

Out[34]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Educ
0	4 1	Yes	Travel_Rarely	1 1 0 2	Sales	1	
1	4 9	No	Travel_Frequently	2 7 9	Research & Development	8	
2	3 7	Yes	Travel_Rarely	1 3 7 3	Research & Development	2	
3	3 3	No	Travel_Frequently	1 3 9 2	Research & Development	3	
4	2 7	No	Travel_Rarely	5 9 1	Research & Development	2	
5	rows × 3 5 columns						

## Import ML Models

## Align Features for Prediction

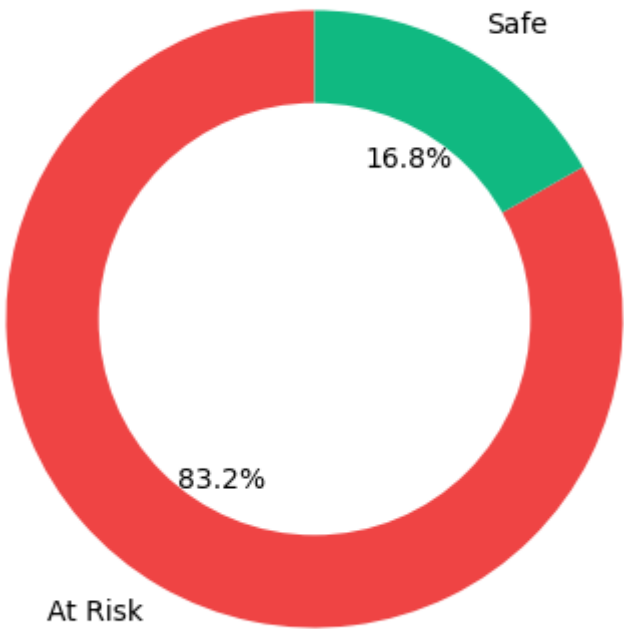
Out[39]:

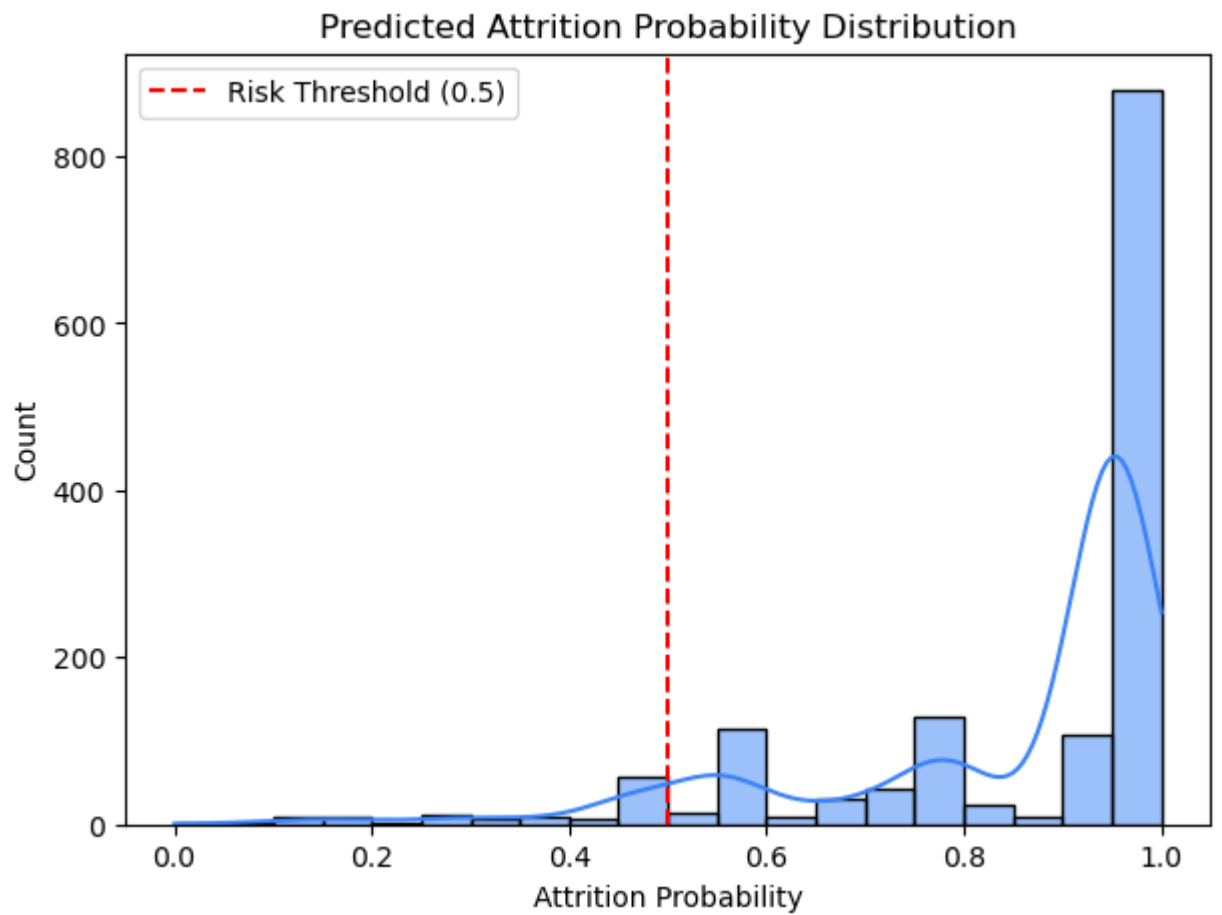
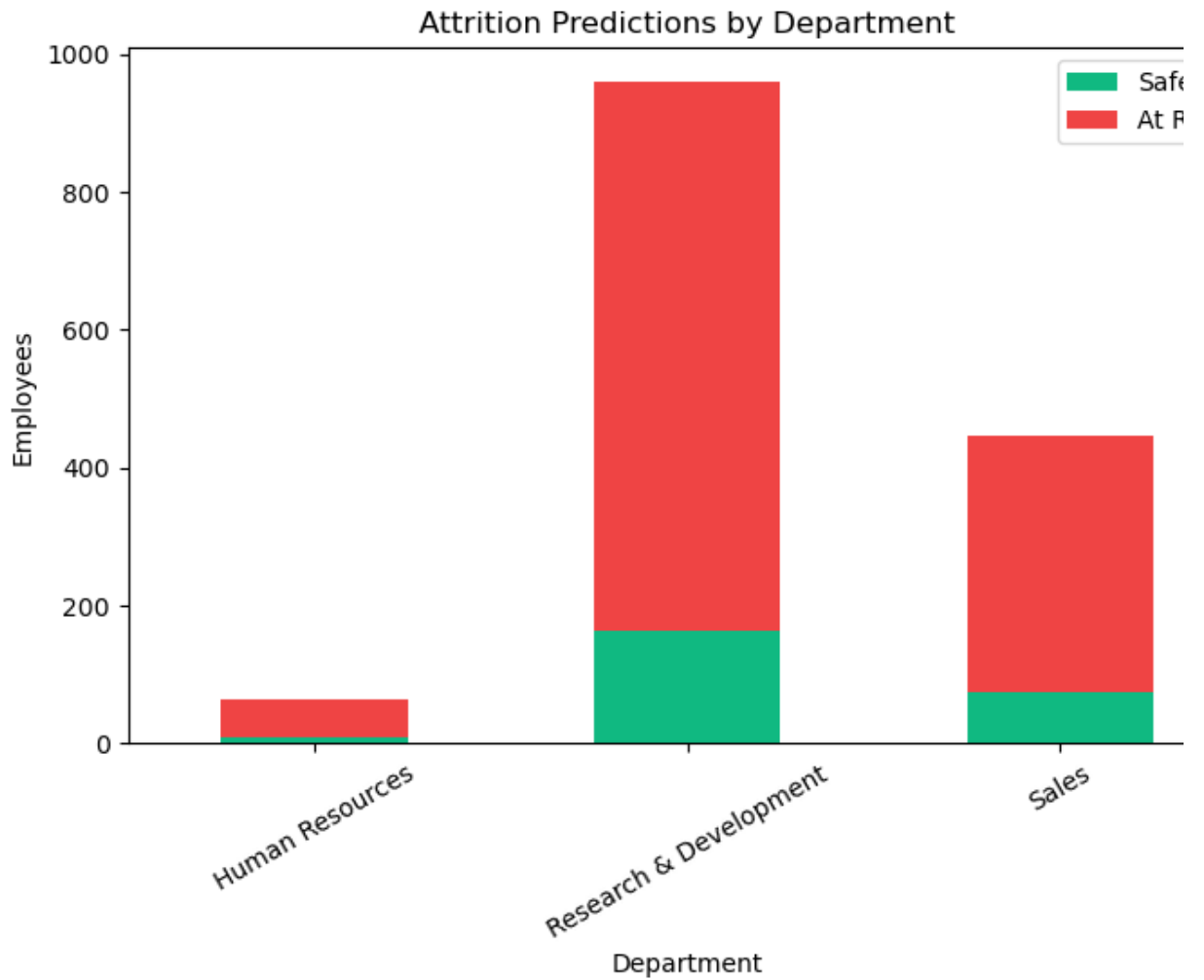
	Age	Department	JobRole	Predicted	Probability
0	4 1	Sales	Sales Executive	1	0 . 9 5 0 6 8 8
1	4 9	Research & Development	Research Scientist	1	0 . 9 4 8 0 4 4
2	3 7	Research & Development	Laboratory Technician	1	0 . 9 5 2 6 3 5
3	3 3	Research & Development	Research Scientist	1	0 . 9 4 8 0 4 4
4	2 7	Research & Development	Laboratory Technician	1	0 . 9 5 0 6 8 8
5	3 2	Research & Development	Laboratory Technician	1	0 . 9 4 8 0 4 4
6	5 9	Research & Development	Laboratory Technician	1	0 . 7 6 3 6 6 1
7	3 0	Research & Development	Laboratory Technician	1	0 . 9 4 8 0 4 4
8	3 8	Research & Development	Manufacturing Director	1	0 . 8 8 5 5 3 1
9	3 6	Research & Development	Healthcare Representative	1	0 . 8 5 2 6 2 1

## Add Visuals - Donut & Dept. Breakdown

⚠ Dropped 'Attrition' column from inference data to avoid leakage.  
🔍 Prediction Debug  
Probability range: 0.0 to 1.0  
Sample stats: [0.46418643 0.7870295 0.95263028 0.95263028 0.96479201]  
✅ Using threshold 0.65 → At Risk: 1223, Safe: 247

At Risk vs Safe (Predictions from SQL data)







This project demonstrated how SQL queries can be seamlessly combined with Machine Learning to run real-time attrition predictions.

## Key Takeaways:

- **Database Integration:** HR data stored in SQL (SQLite) was queried directly inside Python.
- **Leakage Prevention:** Attrition labels were properly excluded from inference data.
- **Predictions:** Logistic/XGBoost models predicted attrition risk per employee.
- **Visuals:**
  - Donut Chart → At Risk vs Safe employees.
  - Department-level bar chart → attrition distribution across functions.
  - Probability distribution → highlights prediction spread & threshold sensitivity.
- **Threshold Optimization:** Added adaptive cutoffs to balance risk prediction and reduce fals

## Artifacts Produced:

- `hr_dataset.db` → SQLite database with IBM HR data (table = employees).
- SQL utility module → `sql_utils.py` for safe querying & reusable functions.
- Visual charts (saved in `/charts/`):
  - `donut_chart.png`
  - `department_attrition.png`
  - `probability_distribution.png`
- Notebook → with integrated SQL + ML pipeline.

## Business Value:

- HR leaders can **query directly** for attrition insights without touching Python code.
  - Predictive analytics embedded into HRIS-like SQL workflows.
  - Foundation for **real dashboards** (Streamlit / BI tools) where HR managers can pull SQL → predictions → export reports.
-