HR Attrition Analytics (Excel-SQL-Python-Power BI)

Author: Fikolawole Oluwadare

Dataset: IBM HR Analytics Employee Attrition & Performance

Executive Summary

Goal: Understand drivers of employee attrition and build a simple, explainable model to study and predict attrition rates.

Outcome: Identified key drivers (monthly income, age bands, job role, department, work-life balance) and produced a baseline classification model with **87% accuracy** on holdout data, plus a Power BI dashboard for business monitoring.

Business value: Targeted retention actions for roles/departments with high risk of attrition.

Dataset Overview

- ✓ **Source**: Kaggle.com IBM HR Analytics (public)
- ✓ **Rows**: ~1.4K records.
- ✓ **Target**: `Attrition` (Yes/No).
- ✓ **Key fields**: Age, Department, JobRole, MonthlyIncome, JobSatisfaction, EnvironmentSatisfaction, WorkLifeBalance, YearsAtCompany, Overtime, DistanceFromHome, Education, MaritalStatus, etc.

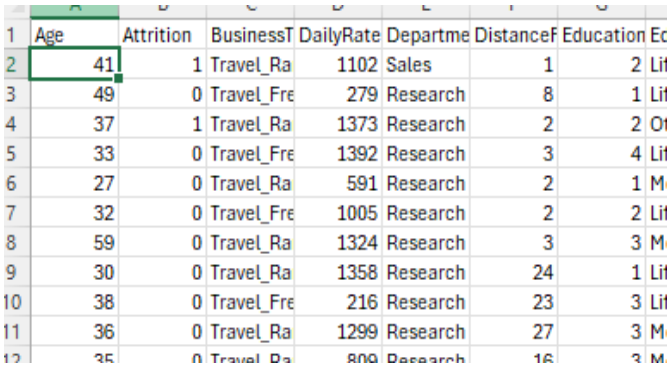
Tools & Workflow

Software: Excel → SQL Server → Python (scikit-learn) → Power BI

1. **Data Cleaning and shaping:**

- I studied the data set to know what business insight I can get from the given columns. I ensure consistent data types, removed noise and standardized the targets
- I removed the irrelevant columns to make my project well clean and only focused on the main goals
- I converted the **Attrition** column from Yes/No to 1/0. Excel formular **=IF(C2="Yes", 1,)**

- Imported cleaned Csv file to SQL for business queries



	Age	Attrition	BusinessT	DailyRate	Departme	DistanceF	Education	Ed
2	41	1	Travel_Ra	1102	Sales	1	2	Lif
3	49	0	Travel_Fre	279	Research	8	1	Lif
4	37	1	Travel_Ra	1373	Research	2	2	Ot
5	33	0	Travel_Fre	1392	Research	3	4	Lif
6	27	0	Travel_Ra	591	Research	2	1	M
7	32	0	Travel_Fre	1005	Research	2	2	Lif
8	59	0	Travel_Ra	1324	Research	3	3	M
9	30	0	Travel_Ra	1358	Research	24	1	Lif
10	38	0	Travel_Fre	216	Research	23	3	Lif
11	36	0	Travel_Ra	1299	Research	27	3	M
12	35	0	Travel_Ra	809	Research	16	3	M

Fig 1. Attrition 1/0

SQL Server (Exploratory & Business Queries)

Loaded the cleaned CSV file to SQL Server, I looked at the loaded table before I began my business queries.

```sql
SELECT TOP 5 * FROM hr_data;

--What is the total number of employees?
SELECT COUNT(*) AS [Total Number of Employees]
FROM hr_data;

--How many employees left the company?
select count(*) as [Employees that Left]
from hr_data where Attrition = '1';

--Attrition rate by department?
SELECT Department, COUNT(*) AS Total,
SUM(CASE WHEN Attrition = '1' THEN 1 ELSE 0 END) AS [Employees that Left]
FROM hr_data GROUP BY Department;

--Average age of employees who left vs stayed
SELECT Attrition, AVG(Age) AS AvgAge
FROM hr_data GROUP BY Attrition;

--Job roles with the highest attrition?
SELECT JobRole, COUNT(*) AS Leavers
 FROM hr_data WHERE Attrition = '1' GROUP BY JobRole  ORDER BY Leavers DESC;

 --Average income by job role
 SELECT JobRole, AVG(MonthlyIncome) AS AvgIncome
 FROM hr_data GROUP BY JobRole;

 --Attrition by overtime status
 SELECT OverTime, COUNT(*) AS Total, SUM(CASE WHEN Attrition = '1' THEN 1 ELSE 0 END) AS Leavers
 FROM hr_data GROUP BY OverTime;

 --Satisfaction score comparison: left vs stayed
 SELECT Attrition, AVG(JobSatisfaction) AS AvgSatisfaction
 FROM hr_data GROUP BY Attrition;

 --Top 3 departments with most employees
  SELECT TOP 3 Department, COUNT(*) AS Total
  FROM hr_data GROUP BY Department ORDER BY Total DESC;

  --Correlation check: YearsAtCompany vs Attrition
  SELECT Attrition, AVG(YearsAtCompany) AS AvgTenure
  FROM hr_data GROUP BY Attrition;
```

Fig 2.

Fig 2 shows the queries that were ran using the **COUNT, AVERAGE, SUM** and more commands to determine the main causes of attrition. These outcomes are good business insight for the employer to know where and how to improve satisfaction of the employees by looking into the pay, job roles, overtime and length of stay in the company. This would be very useful for the Human resources department.

Python (Modeling & Explainability)

The goal of this stage was to use **Python** to build predictive models for employee attrition, based on the cleaned and prepared dataset from earlier steps (Excel cleaning + SQL querying). The models aim to help HR teams identify which employees are most likely to leave and what factors drive attrition.

```
import pandas as pd
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, roc_auc_score, roc_curve, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns


df = pd.read_csv("IBM _HR_Cleaned.csv")
```

Fig 3. List of imported libraries used for the python analysis of this project.

- I Imported the dataset into Python using Pandas.
- I Split data into features (X) and target (y) where Attrition was the target variable.
- I Applied train-test split (80/20) for unbiased model evaluation.

I used **logistic regression** as baseline model and **random forest** to capture nonlinear relationship and feature importance, I also used **SHAP** explainability for more transparency.

Models were trained on the training data set and predictions were made on the test dataset and performance was accessed using multiple metrics like **accuracy, precision, recall, F1-score and confusion matrix.**

RESULTS

Logistic regression.

```
=== Logistic Regression ===
              precision    recall  f1-score   support

           0       0.92      0.79      0.85       247
           1       0.36      0.64      0.46        47

    accuracy                           0.76       294
   macro avg       0.64      0.71      0.65       294
weighted avg       0.83      0.76      0.79       294

ROC-AUC: 0.8072185373417177
```

Fig 4.

The Logistic Regression model achieved **76% accuracy** and a **ROC-AUC of 0.81**, showing good overall discrimination. With an attrition rate of **16%** 47 of 294 employees, the model performed strongly for employees who stayed with **precision** of 0.92 and **F1-score** of 0.85 but less effectively for those who left **precision** of 0.36 and a **recall** of 0.64). Overall, it predicts retention well but is less precise on attrition, so I tried a more advanced model like **Random Forest** and see how it works.

RANDOM FOREST

```
=== Random Forest ===
              precision    recall  f1-score   support

           0       0.85      0.98      0.91       247
           1       0.45      0.11      0.17        47

    accuracy                           0.84       294
   macro avg       0.65      0.54      0.54       294
weighted avg       0.79      0.84      0.79       294

ROC-AUC: 0.7729778620036178
```

Fig 5.

The **Random Forest** model achieved **84%** accuracy and a **ROC-AUC** of **0.77**. With an attrition rate of **16%** (47 of 294 employees), it performed very well on employees who stayed with a **precision** of 0.85, **recall** of 0.98 and a **F1-score** of 0.91, but struggled with employees who left with a **precision** of 0.45, **recall** of 0.11and a **F1-score** of 0.17. Overall, the model is highly reliable at identifying retention but performs poorly in detecting attrition just like **logistic regression.**

```
Confusion Matrix:
 [[253   2]
 [ 36   3]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.99      0.93       255
           1       0.60      0.08      0.14        39

    accuracy                           0.87       294
   macro avg       0.74      0.53      0.53       294
weighted avg       0.84      0.87      0.82       294
```

Fig 6.

Using the confusion matrix with the **random forest** model, the model accurately identifies employees who will stay, with 253 true negatives and only 2 false positives. However, it struggles to detect employees who will leave, with only 3 true positives and 36 false negatives. This results in high recall for "**No Attrition**" (0.99) but very low recall for "**Attrition**" (0.08). Overall accuracy is 0.87, but the predictions are heavily skewed toward staying employees. While the model performs well for general workforce retention predictions, it may need further tuning or additional features to better identify at-risk employees.

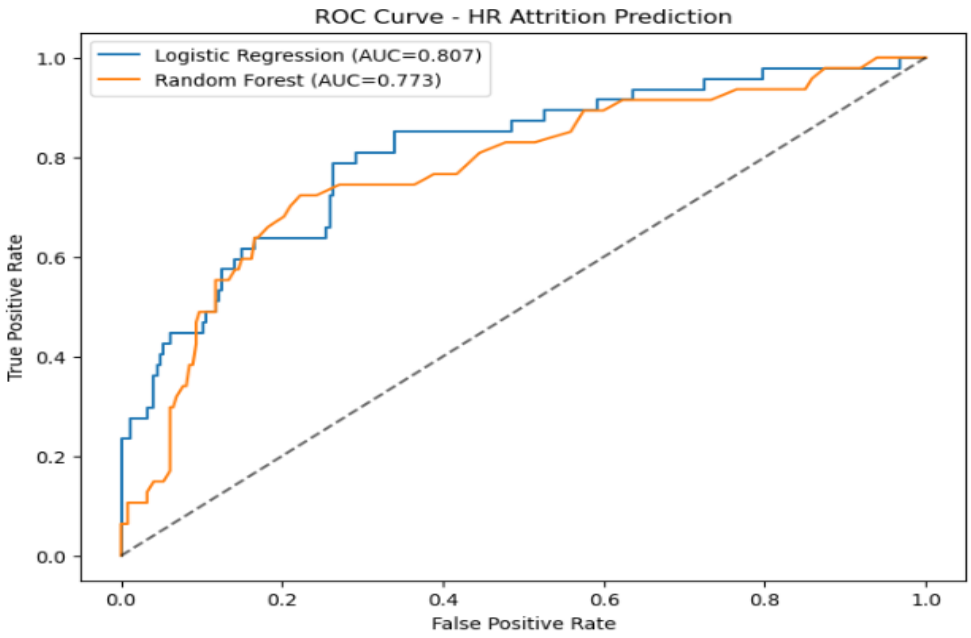


Fig 7.

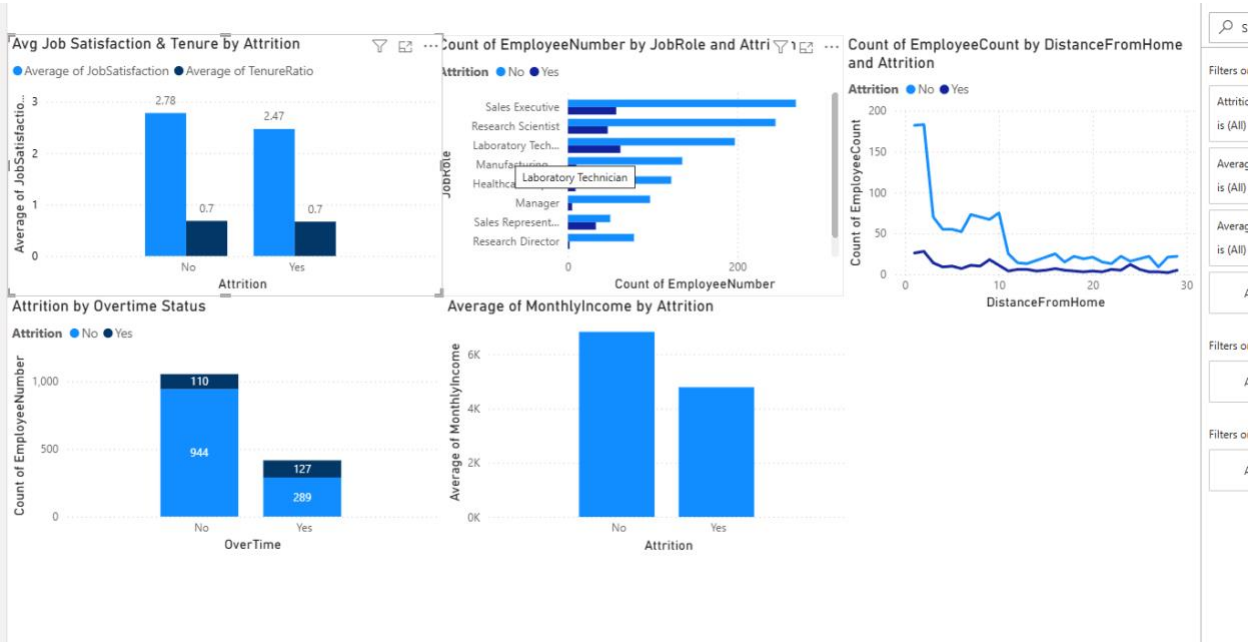
The ROC curve compares the ability of **Logistic Regression** and **Random Forest** to distinguish between employees who stay and those who leave.

- **Logistic Regression** achieved an AUC of **0.807**, indicating stronger predictive discrimination.
- **Random Forest** achieved an AUC of **0.773**, performing slightly lower but still well above random guessing (AUC = 0.5).

Interpretation:
Both models perform reasonably well, but Logistic Regression is more effective at ranking employees by attrition risk, as shown by its higher AUC and steeper curve. Random Forest, while slightly weaker in this metric, remains useful for identifying feature importance and understanding the key drivers of attrition.

Conclusion:
Logistic Regression is the stronger model for predicting attrition, while Random Forest complements it by providing valuable insights into the factors influencing employee turnover.

Power BI Report – HR Attrition Analysis



This Power BI dashboard explores key factors influencing employee attrition, providing actionable insights into workforce dynamics.

Key Findings

- **Job Satisfaction & Tenure**: Employees who left reported **lower average job satisfaction (2.47 vs 2.78)** compared to those who stayed.

  Tenure ratio was also slightly lower for employees who left (0.7), suggesting that shorter tenure may contribute to attrition.

- **Job Role**: Attrition is most common among **Sales Executives, Research Scientists, and Laboratory Technicians**, highlighting roles with higher turnover risk.
- **Distance from Home:** Employees living farther from the workplace show higher attrition counts, indicating that **commuting distance** may be a factor in turnover decisions.
- **Overtime**: A clear pattern emerges among employees working overtime, **127 left compared to 289 who stayed**, showing that overtime work is strongly linked with higher attrition.
- **Monthly Income**: Employees who left had a **lower average monthly income ~$4,500** compared to those who stayed **~$6,000+**.

  This suggests compensation is a significant driver of attrition.

Conclusion

The Power BI analysis shows that attrition is influenced by multiple factors, including **low job satisfaction, high overtime workload, longer commuting distance, and lower income**. Job roles such as Sales Executives and Research Scientists are particularly vulnerable. These insights can help HR teams focus on targeted retention strategies, such as improving compensation, managing overtime, and supporting employees in high-risk roles.

Business Insight

Attrition is highest among employees with **lower job satisfaction, longer commutes, lower income, and heavy overtime workloads**. Certain roles such as **Sales Executives and Research Scientists** are most at risk. Addressing these areas through better compensation, workload balance, and targeted retention initiatives could significantly reduce turnover.

1. **Work-life balance is protective**: Higher WLB scores strongly correlate with retention. Managers should prioritize flexible scheduling and wellness initiatives to improve WLB scores in departments with high attrition.
2. **Tenure matters**: Attrition risk is highest in the first 12–18 months. Early engagement, mentorship, and recognition programs could increase retention during this critical window.

3. **Younger workforce mobility**: Employees under 30 show a higher likelihood of leaving, suggesting the need for stronger career progression pathways and targeted development plans.
4. **Commute burden**: Long distance from home is linked to attrition, hinting at the value of hybrid or remote work policies.

Overall, the combination of **targeted retention programs** and **policy level interventions** has the potential to materially lower turnover rates and associated costs.

Recommendations

- **Improve Compensation**: Review pays structures for lower-income roles, particularly Sales Executives and Research Scientists, to reduce attrition linked to salary gaps.
- **Manage Overtime**: Implement workload balancing and overtime limits to prevent burnout and turnover.
- **Enhance Job Satisfaction**: Launch engagement programs, career development plans, and recognition initiatives to improve employee satisfaction.
- **Address Commute Challenges**: Explore flexible work options, hybrid schedules, or relocation support for employees with long commutes.
- **Target High-Risk Roles**: Prioritize retention strategies for job functions with higher attrition rates, tailoring support to their specific challenges.