IBM HR Analytics — Employee Attrition & Performance

Submitted By: Monu Ramkesh Gupta

Project: IBM HR Analytics — Employee Attrition & Performance Report

Course / Internship: Data Analyst Internship

Executive Summary

This report presents an end-to-end analysis of the IBM HR dataset focusing on **employee attrition and related performance factors**. The goals were to: (1) explore patterns and drivers of attrition, (2) build predictive insights using machine learning, and (3) deliver an interactive Power BI dashboard for business stakeholders.

Key outcomes: - Dataset: ~1,470 employees; Attrition = 237 (16%). - Top drivers identified by Random Forest (top features listed in results section). - Interactive Power BI dashboard with KPI cards, demographic & compensation charts, and model-based feature importance.

Table of Contents

---

1. Introduction

Employee attrition is a critical HR metric with direct impact on costs and productivity. This project analyzes IBM HR data to find who is leaving and why, and suggests focus areas for retention.

2. Objective

- Calculate attrition metrics and key KPIs.
- Explore relationships between attrition and demographic / workplace factors (gender, job role, income, distance from home, job satisfaction, etc.).
- Build simple predictive models and extract feature importance.
- Deliver an interactive, business-ready Power BI dashboard.

3. Dataset Description

- Source: IBM HR Analytics dataset (commonly used HR attrition dataset).
- Rows: ~1,470 employees
- Target: `Attrition` (Yes / No)

Sample columns used: - EmployeeID (unique identifier) - Age, Gender, Department, JobRole - MonthlyIncome, DistanceFromHome - JobSatisfaction, PerformanceRating - Attrition (Yes/No)

4. Data Preparation

Steps performed: 1. Load data into pandas DataFrame. 2. Check missing values and datatypes: `df.info()` and `df.isnull().sum()`. 3. Handle missing values where necessary (drop or impute). 4. Convert categorical columns to appropriate types (e.g., `astype('category')`). 5. Create derived features: `Age Group` (bins), `Income Band` (optional). 6. Split data into train/test for modeling (e.g., 80% train, 20% test).

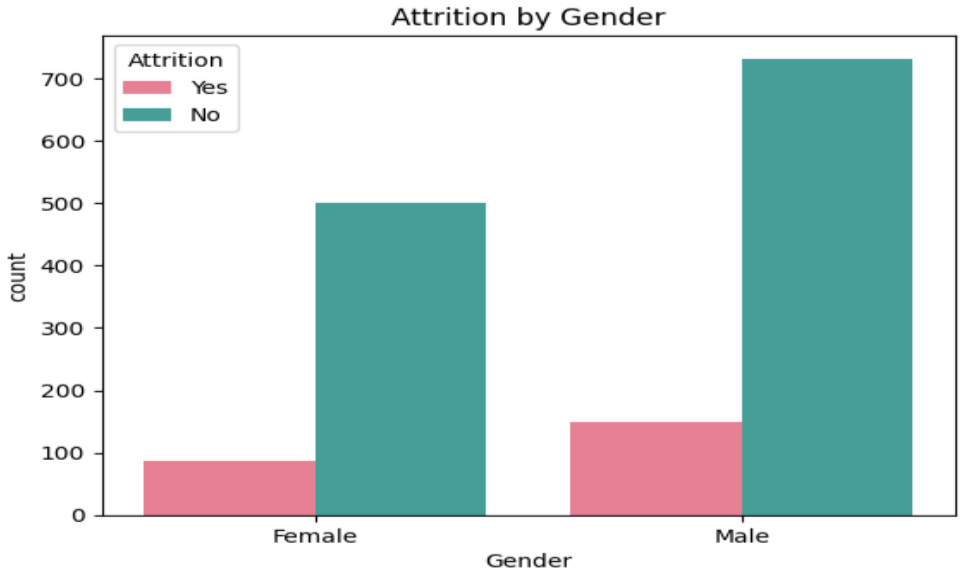
Example code snippets (Python):

```
# Load
import pandas as pd
df = pd.read_csv('IBM_Dataset.csv')

print(df.shape)
print(df.info())

# Age group example
def age_group(age):
    if age <= 30: return '20-30'
    if age <= 40: return '31-40'
    if age <= 50: return '41-50'
    return '51+'

df['AgeGroup'] = df['Age'].apply(age_group)
```

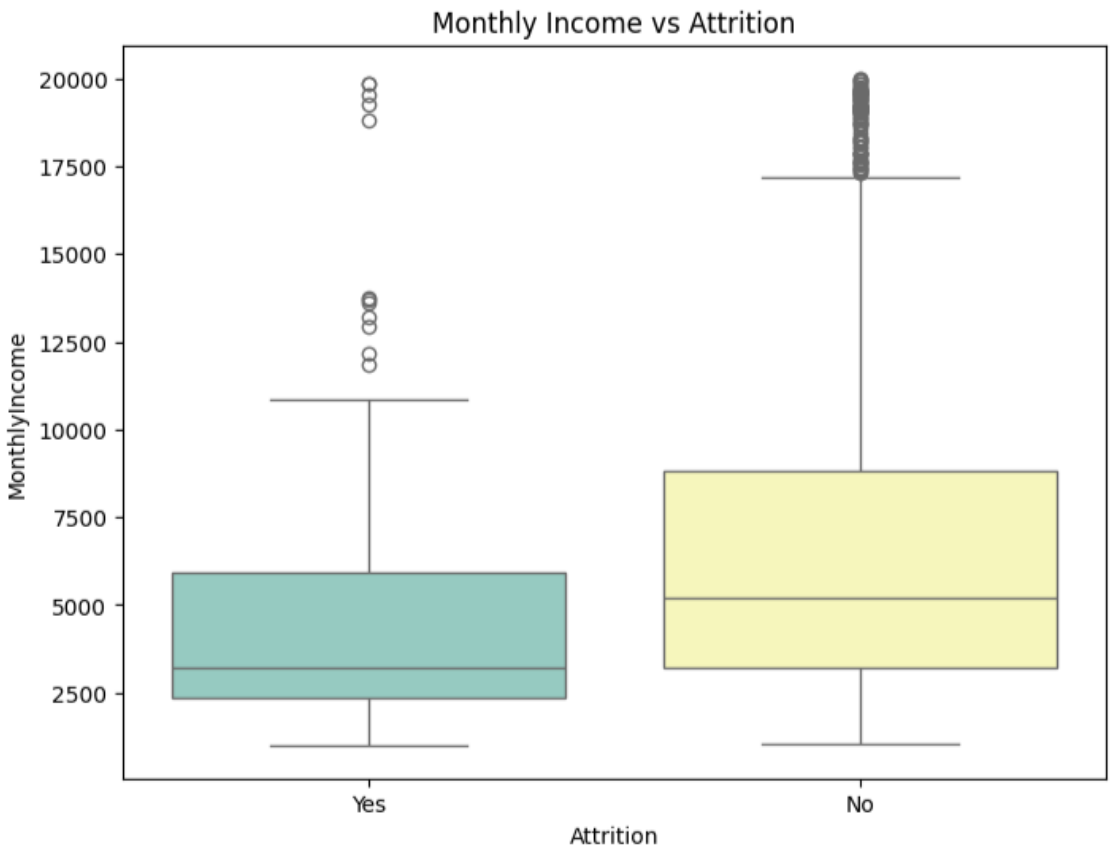
---

5. Exploratory Data Analysis (EDA)

Main plots created — Attrition by Gender.

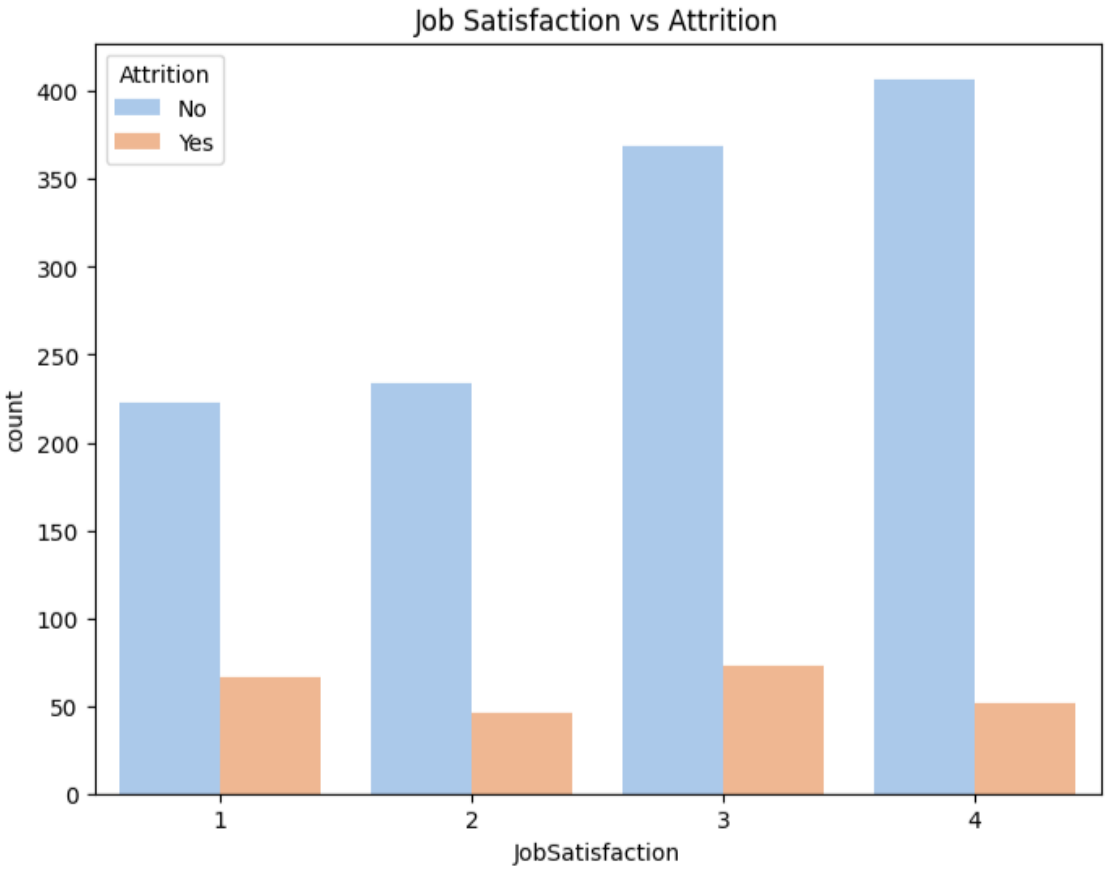


- `sns.boxplot(x='Attrition', y='MonthlyIncome')` — Income distribution by Attrition.



- `sns.barplot(x='JobRole', y='DistanceFromHome', hue='Attrition')` — Distance from Home by Job Role & Attrition.

— Job Satisfaction vs Attrition.



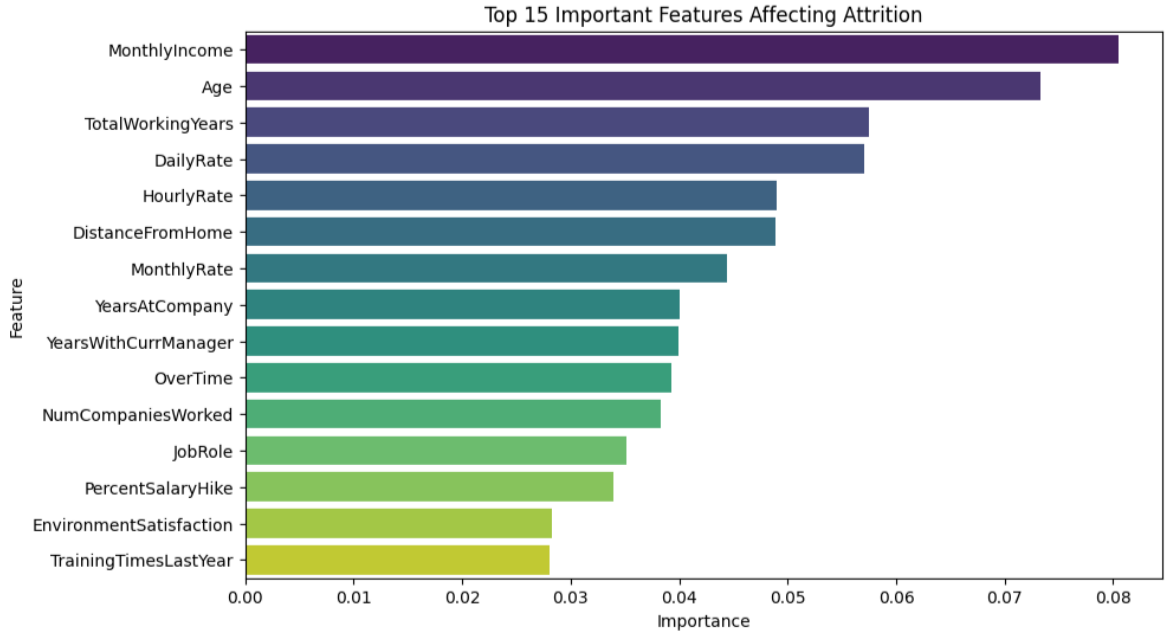
6. Modeling

Models tried: Logistic Regression, Decision Tree, Random Forest.

Evaluation metrics used: Accuracy, Precision, Recall, F1-score, Confusion Matrix.

Confusion matrices: - Logistic Regression - Decision Tree - Random Forest

Feature importance (Random Forest): - Extracted with `rf_model.feature_importances_` and placed into DataFrame sorted descending.



Code snippet (feature importance):

```
importances = rf_model.feature_importances_


feat_imp = pd.DataFrame({
    'Feature': X.columns,
    'Importance': importances
}).sort_values(by="Importance", ascending=False)


# Show top 10 features
print("\nTop 10 Important Features:\n", feat_imp.head(10))
```

```
# Plot Top 15

plt.figure(figsize=(10,6))

sns.barplot(x="Importance", y="Feature", data=feat_imp.head(15),
palette="viridis")

plt.title("Top 15 Important Features Affecting Attrition")

plt.show()
```

---

7. Dashboard & Visualizations

---

8. Results & Insights

Overall attrition rate = **16%** (237/1470). - Job Roles with highest attrition: *Sales Executive, Laboratory Technician, etc.* - Employees with **low JobSatisfaction** and **high OverTime** have higher attrition. - Feature importance shows **MonthlyIncome**, **Age**, **JobRole**, **DistanceFromHome** as top contributors.

Recommendations for HR: - Target retention programs for high-risk job roles. - Review compensation / benefits for roles with high attrition. - Investigate workload and overtime policies. - Conduct targeted employee engagement surveys for low satisfaction groups.

---

9. Conclusion

Summarize the analysis, key findings, and business recommendations. Emphasize how the dashboard enables HR to quickly identify at-risk segments and take data-driven actions.

---

10. References

- IBM HR Analytics dataset (UM)
- pandas / seaborn / scikit-learn documentation
- Power BI documentation
- YouTube-tutorials, or videos
- Google
