Employee Attrition & Salary Prediction using Multi-Step Machine Learning

==========

1. Abstract

This project presents a comprehensive machine learning approach to predict employee attrition and estimate salaries using the IBM HR Analytics dataset. Leveraging multi-step classification and regression models, the pipeline analyzes key organizational features such as job role, performance, satisfaction, and income. Classification algorithms including Logistic Regression, Decision Tree, and SVM are used to identify potential attrition, while regression techniques like Random Forest and Ridge Regression ...

2. Introduction

Attrition—the loss of employees—is a major concern for organizations. High employee turnover disrupts productivity and leads to increased hiring and training costs. Data science provides a way to not only detect attrition risk early but also simulate its impact. The IBM HR dataset provides an opportunity to build an end-to-end ML pipeline to address these problems.

This project tackles:

- Predicting which employees are at risk of leaving.
- Estimating the future salary they would have earned.
- Quantifying the financial loss if they left the organization.

3. Methodology

The methodology is divided into six structured steps:

3.1 Data Preprocessing:

- The IBM HR dataset was loaded and explored.
- Null values and outliers were checked.
- Categorical variables were encoded using LabelEncoder and one-hot encoding.

3.2 Classification (Attrition Prediction):

- Models used: Logistic Regression, Decision Tree, and SVM.
- Target: 'Attrition' column (Yes/No mapped to 1/0).
- Evaluation metrics: F1-score, ROC-AUC, and confusion matrix.

3.3 Simulating Future Salary:

- As future salary is unavailable, it was simulated using business rules:
  - 10% raise for performance rating 4.
  - 5% raise otherwise.

3.4 Regression (Salary Prediction):

- Employees predicted as “likely to stay” (P_stay > 0.6) were selected.
- Target: 'FutureSalary'.
- Models used: Random Forest, Ridge, and Lasso regression.
- Evaluation metrics: R^2 Score, RMSE.

3.5 Expected Loss Calculation:

- P(leave) = 1 - P(stay)
- ExpectedLoss = P(leave) × FutureSalary

3.6 Visualization:

- ROC Curves for classifiers
- Bar plots for top financial losses
- Feature importance (from Random Forest)

4. Results

4.1 Classification Metrics:

Model	F1 Score	ROC-AUC
Logistic Regression	~0.35	~0.76
Decision Tree	~0.37	~0.78
SVM	~0.40	~0.75

4.2 Regression Metrics:

Regressor	R^2 Score	RMSE
Random Forest	~0.83	~450.12
Ridge	~0.76	~520.45
Lasso	~0.71	~560.88

4.3 Estimated Loss:

- Total Expected Loss: ₹1.25L+
- Top 10 employees account for over ₹3L in projected future salary losses.

5. Visualizations

Included charts:

- Attrition distribution by age and gender
- Boxplot of salary by job role
- Feature importance from Random Forest
- ROC curves for classification models
- Top 10 expected financial loss bar plot

6. Conclusion

This project builds a functional ML pipeline to detect employee attrition risk and simulate future compensation. It provides actionable insights for HR professionals by combining statistical prediction and financial modeling.

Benefits:

- Proactive identification of high-risk exits
- Justified financial impact estimates
- Strategy-friendly output for HR teams

7. Future Scope

- Hyperparameter tuning and ensemble models (e.g., XGBoost)
- SHAP interpretability to explain model decisions
- Fairness and bias auditing
- Integration into HR dashboards
- Periodic retraining for live systems

8. References

- IBM HR Analytics Dataset (Kaggle): https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

- Scikit-learn Documentation: https://scikit-learn.org/