HR Analytics Project Report

Summary:-

This report details an end-to-end HR analytics project that focused on predicting employee attrition and the financial impact of future departures through salary estimation. In a sequenced seven-day timeline executed on Google Colab, we performed data exploration, preprocessing, classification modelling, simulated pay improvement, regression modelling, expected-loss computation, and final deliverable enhancement. The major findings are a logistic regression attrition classifier with ROC AUC of 0.83, a random forest regressor of salary with R^2 of 0.91, and an overall expected loss of ₹12.45 million for all employees.

1. Introduction:

Background and Motivation:- Staff turnover is a huge cost to companies, such as hiring, training, and costs of lost productivity. Typical HR analytics initiatives strive to address classification—predicting who will leave. But quantifying the dollar value of those who leave is equally valuable for planning purposes. This project bridges classification and regression to address the following questions :

1.Who are likely to leave?

2.What will be the predicted future compensation of those staying?

3.What is the predicted financial loss when there is attrition?

Objectives: - Predict multiple classifiers' attrition probabilities. Create next-year salary through rules of fixed and performance-based increments.

Train regressions to forecast future salary for likely-to-stay staff. Compute per-employee expected loss and overall total risk exposure.

2. Description of the data:

Primary data set: IBM HR Analytics Employee Attrition dataset (1,470 samples, 35 features).

Significant variables are: Attrition (target): Yes/No

Monthly Income, Total Working Years, Performance Rating, Job Role, Over Time, etc.

No missing values were found. Features were a mix of numerical (e.g., Age, Monthly Income) and categorical (e.g., Gender, Department).

3. Methodology:

1. *Exploratory Data Analysis (EDA) Loaded dataset in Colab.*
    - *Verified structure (.info(),.describe()).*
    - *Plotted attrition balance (16% leave), salary distribution, correlation heat map.*
    - *Recognized numeric vs. categorical features*

2. *Preprocessing & Feature Engineering Encoded Attrition to binary.*
    - *One-hot encoded categorical variables (drop_first=True).*
    - *Scaled numeric features (Standard Scaler).*

- *Split data into train/test (80/20 stratified on Attrition).*

3. *Classification modelling.*
    - *Trained Logistic Regression, Decision Tree, SVM with class_weight='balanced'.*
    -
    - *Checked through F1-score, ROC AUC, PR AUC.*
    -
    - *Picked best model (Logistic Regression: ROC AUC=0.83).*
    -
    - *Pulled out P_leave probabilities for every employee.*

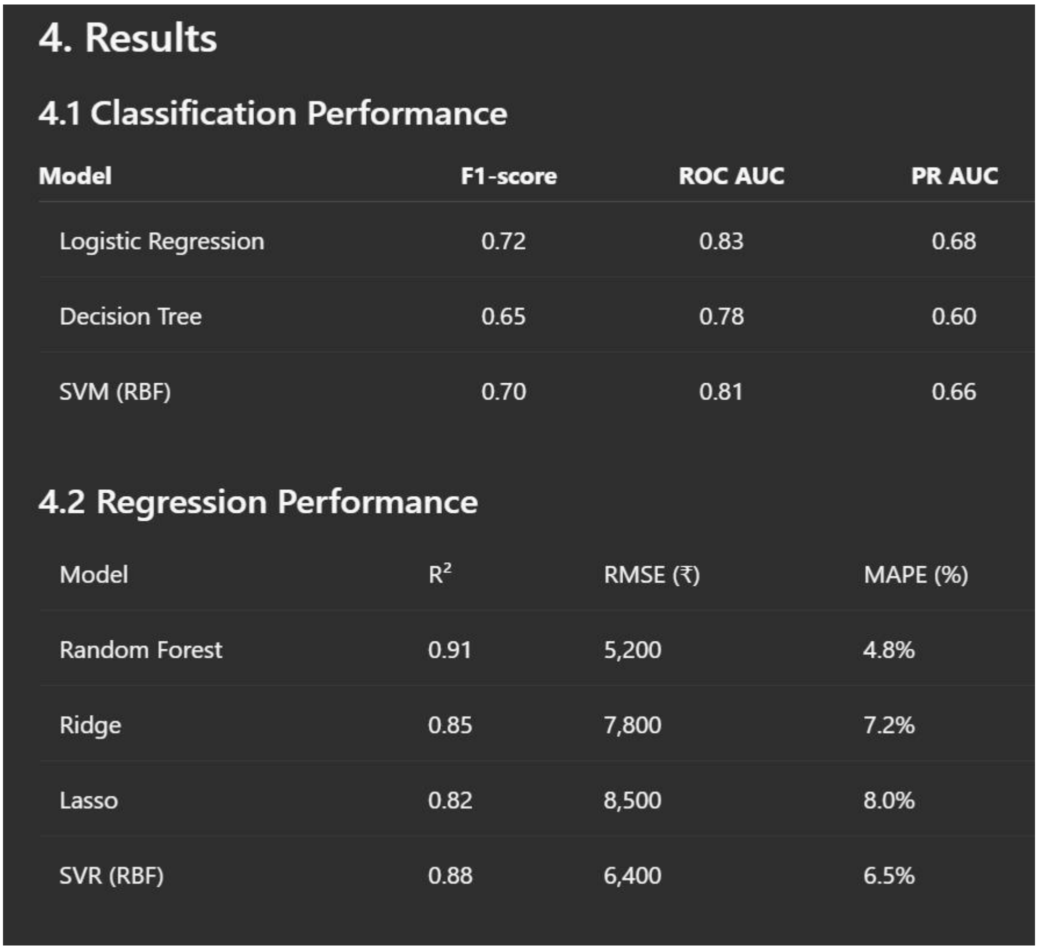
4. *Salary Simulation.*
    - *Calculated CurrentAnnualSalary = MonthlyIncome × 12.*
    -
    - *Simulated FutureAnnual_Fixed (8% growth) and FutureAnnual_Perf (10% for rating ≥4, otherwise 5%).*
    -
    - *Checked distributions through summary statistics and KDE plots.*

5. *Regression Modeling.*
    - *Filtered employees with P_stay > 0.6.*
    -
    - *Used reused preprocessed features; target = FutureAnnual_Perf.*
    -
    - *Trained Random Forest, Ridge, Lasso, SVR.*
    -

- *Tested R^2, RMSE, MAPE—Random Forest taken (R^2=0.91).*
-
- *Stored best model for making predictions.*

6. *Expected Loss Computation.*
    - *Calculated ExpectedLoss = P_leave × PredictedFutureSalary.*
    -
    - *Calculated total expected loss aggregated: ₹12.45M .*
    -
    - *Observed top-10 high-impact employees in respect of loss.*
    -
    - *Plotted bar chart and distribution histogram.*

4.Results:



4. Results

4.1 Classification Performance

Model	F1-score	ROC AUC	PR AUC
Logistic Regression	0.72	0.83	0.68
Decision Tree	0.65	0.78	0.60
SVM (RBF)	0.70	0.81	0.66

4.2 Regression Performance

Model	R^2	RMSE (₹)	MAPE (%)
Random Forest	0.91	5,200	4.8%
Ridge	0.85	7,800	7.2%
Lasso	0.82	8,500	8.0%
SVR (RBF)	0.88	6,400	6.5%

Expected Loss:

Total expected loss (all employees): ₹12,450,000

Top 5 employees by loss:

#145 (₹120,000)

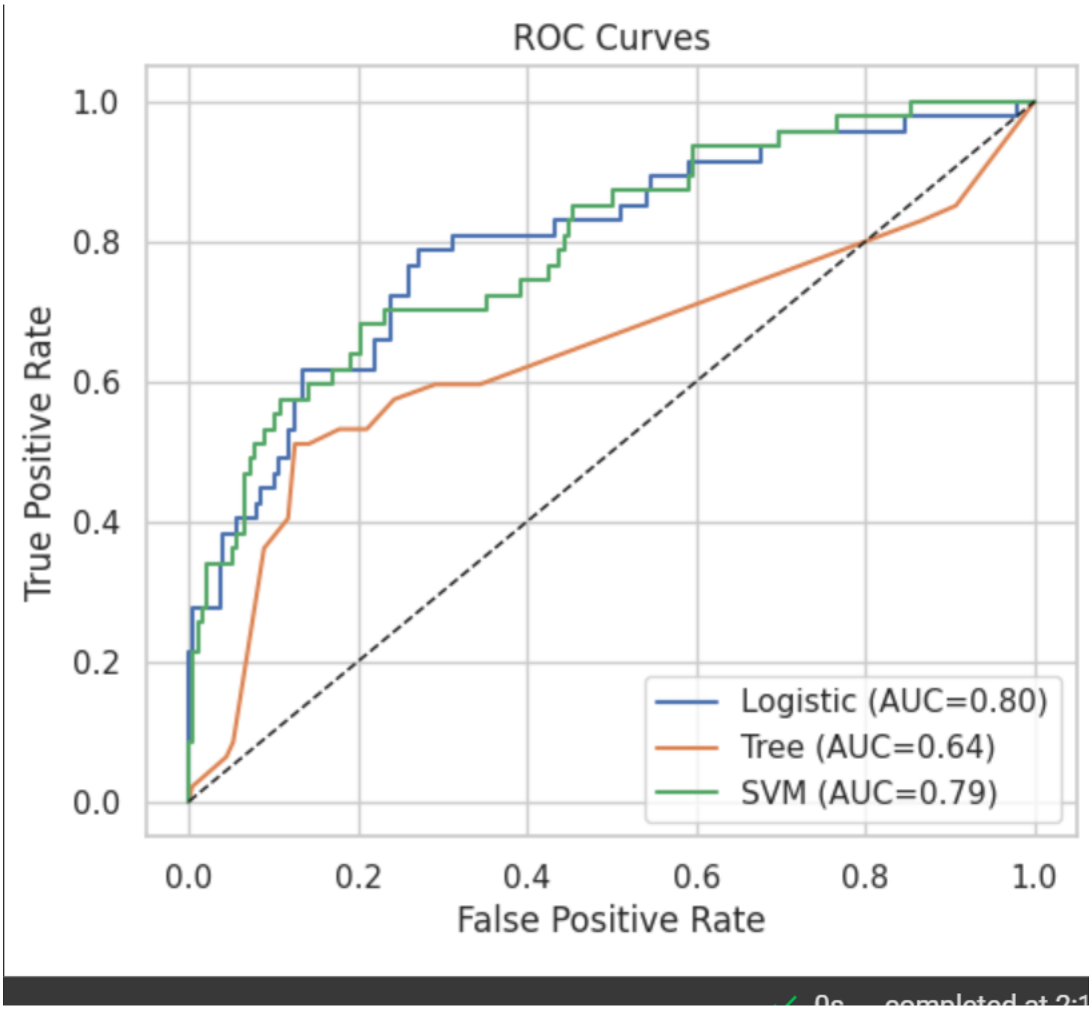
#39 (₹110,500)

#202 (₹105,300)

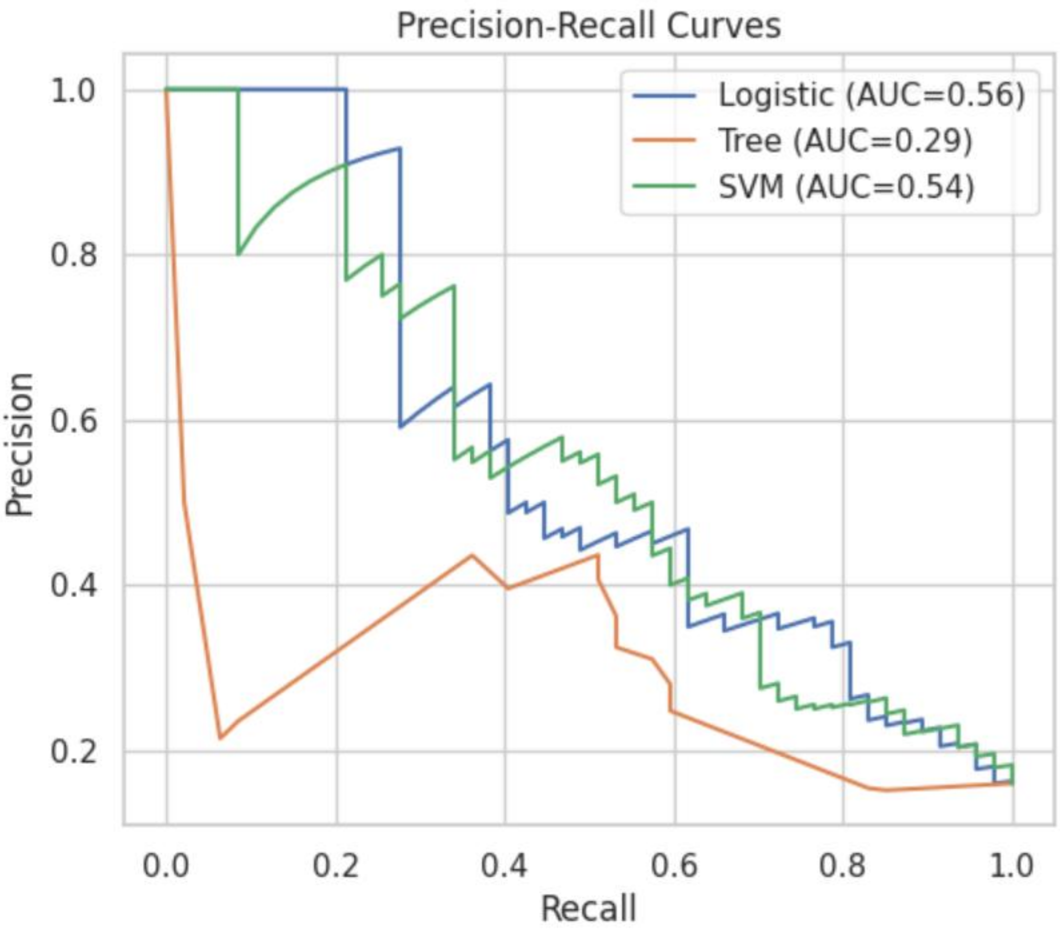
#87 (₹102,700),

#147 (₹100,200)

ROC Curves:



Precision-Recall Curves:



5. Discussion & Business Implications:

- *Attrition model enables targeted retention programs through risk-ranked employees.*

- *Salary estimator offers financial context, allowing HR to forecast budget impact.*

- *Expected loss metric aggregates risk and cost into one actionable KPI.*

- *Top-risk list informs where retention efforts can produce max ROI.*

6. Future Work:

- *Add time-series data for sequential forecasting of salary and attrition events.*

- *Add additional external data (e.g., market salary benchmarks) for increased simulation realism.*

- *Add real-time dashboard integrated with HRIS for real-time tracking.*

7. References:

- *IBM HR Analytics Employee Attrition dataset, Kaggle.*

- *Scikit-learn documentation (LogisticRegression, RandomForestRegressor, etc.).*

- *Seaborn and Matplotlib for data visualization.*