HR Analytics - Predict Employee Attrition

Contribution: Apurba Kuiti

1. Introduction

Employee attrition is a critical issue faced by organizations worldwide. Understanding why employees leave and predicting potential resignations helps HR departments design effective retention strategies. This project applies data analytics and machine learning techniques to identify the key factors that contribute to employee attrition and build predictive models to forecast future attrition trends.

2. Methodology

The project workflow consisted of the following steps:

1. **Data Collection & Cleaning:** The HR dataset was preprocessed using Python (Pandas, NumPy) to handle missing values and encode categorical variables.
2. **Exploratory Data Analysis (EDA):** Conducted in Power BI and Seaborn to visualize department-wise attrition, salary bands, and promotions.
3. **Model Building:** Two classification models were implemented using Scikit-Learn — Logistic Regression and Decision Tree.
4. **Model Evaluation:** Models were evaluated using accuracy, precision, recall, F1-score, ROC AUC, and R² metrics.
5. **Explainability:** SHAP values were used to interpret how features such as Age and DailyRate influence attrition predictions.
6. **Visualization:** Power BI dashboard was created to present insights on attrition patterns and key employee metrics.

3. Results and Analysis

The models produced the following performance metrics:

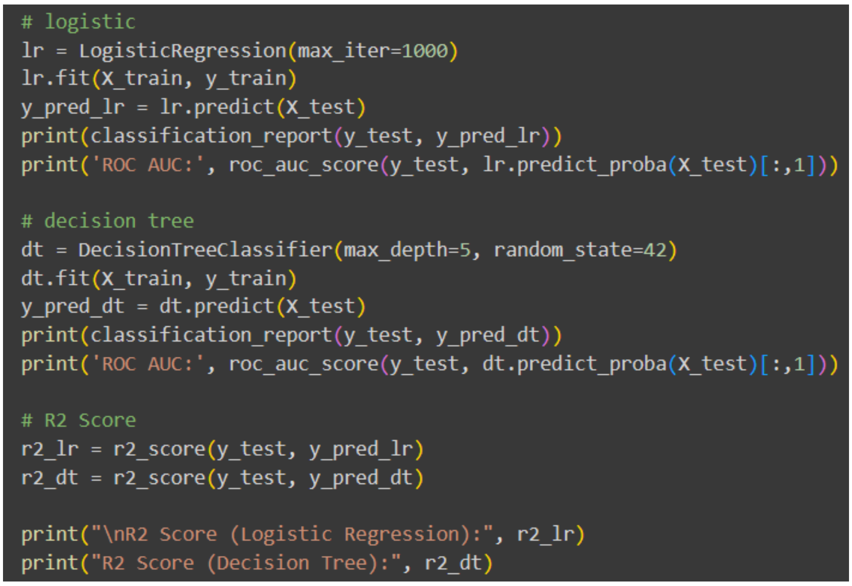
Logistic Regression:

The model achieved strong accuracy and ROC AUC, indicating good distinction between employees who stay and those who leave. However, recall for attrition (class 1) was low, showing that the model missed some actual resignations.

Decision Tree:

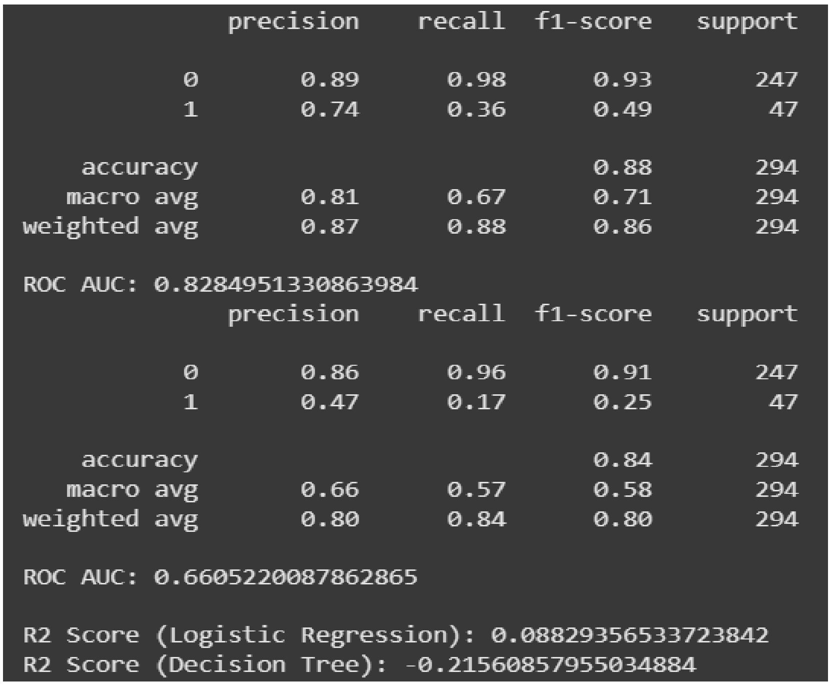
The Decision Tree model slightly underperformed compared to Logistic

Regression and displayed overfitting tendencies.



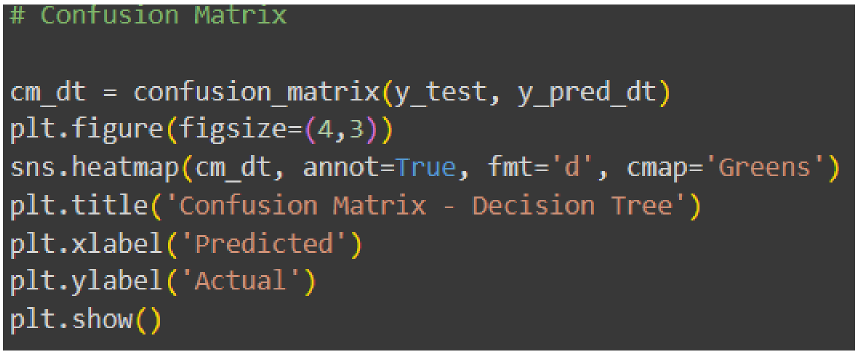
```
# logistic
lr = LogisticRegression(max_iter=1000)
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)
print(classification_report(y_test, y_pred_lr))
print('ROC AUC:', roc_auc_score(y_test, lr.predict_proba(X_test)[:,1]))

# decision tree
dt = DecisionTreeClassifier(max_depth=5, random_state=42)
dt.fit(X_train, y_train)
y_pred_dt = dt.predict(X_test)
print(classification_report(y_test, y_pred_dt))
print('ROC AUC:', roc_auc_score(y_test, dt.predict_proba(X_test)[:,1]))

# R2 Score
r2_lr = r2_score(y_test, y_pred_lr)
r2_dt = r2_score(y_test, y_pred_dt)

print("\nR2 Score (Logistic Regression):", r2_lr)
print("R2 Score (Decision Tree):", r2_dt)
```

Output:

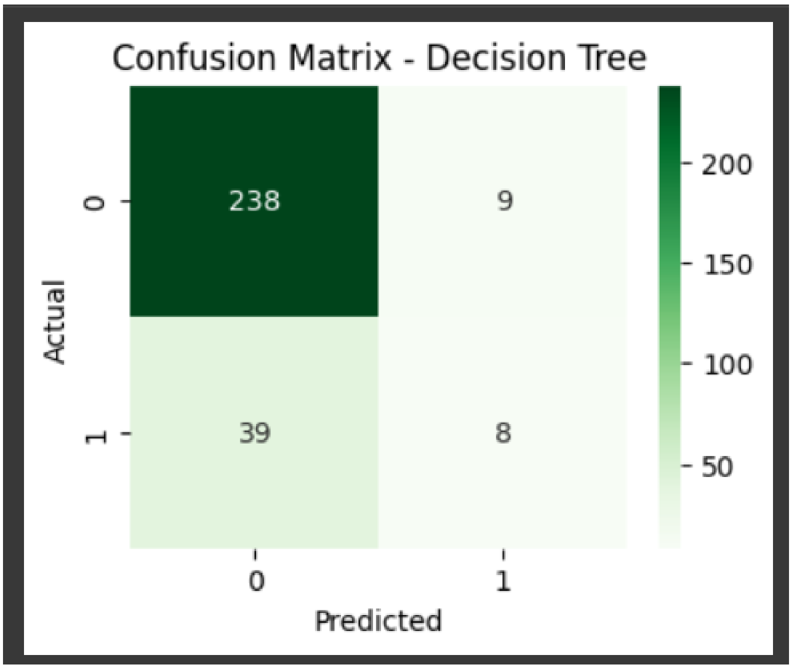


```
              precision    recall  f1-score   support

           0       0.89      0.98      0.93       247
           1       0.74      0.36      0.49        47

    accuracy                           0.88       294
   macro avg       0.81      0.67      0.71       294
weighted avg       0.87      0.88      0.86       294

ROC AUC: 0.8284951330863984
              precision    recall  f1-score   support

           0       0.86      0.96      0.91       247
           1       0.47      0.17      0.25        47

    accuracy                           0.84       294
   macro avg       0.66      0.57      0.58       294
weighted avg       0.80      0.84      0.80       294

ROC AUC: 0.6605220087862865

R2 Score (Logistic Regression): 0.08829356533723842
R2 Score (Decision Tree): -0.21560857955034884
```

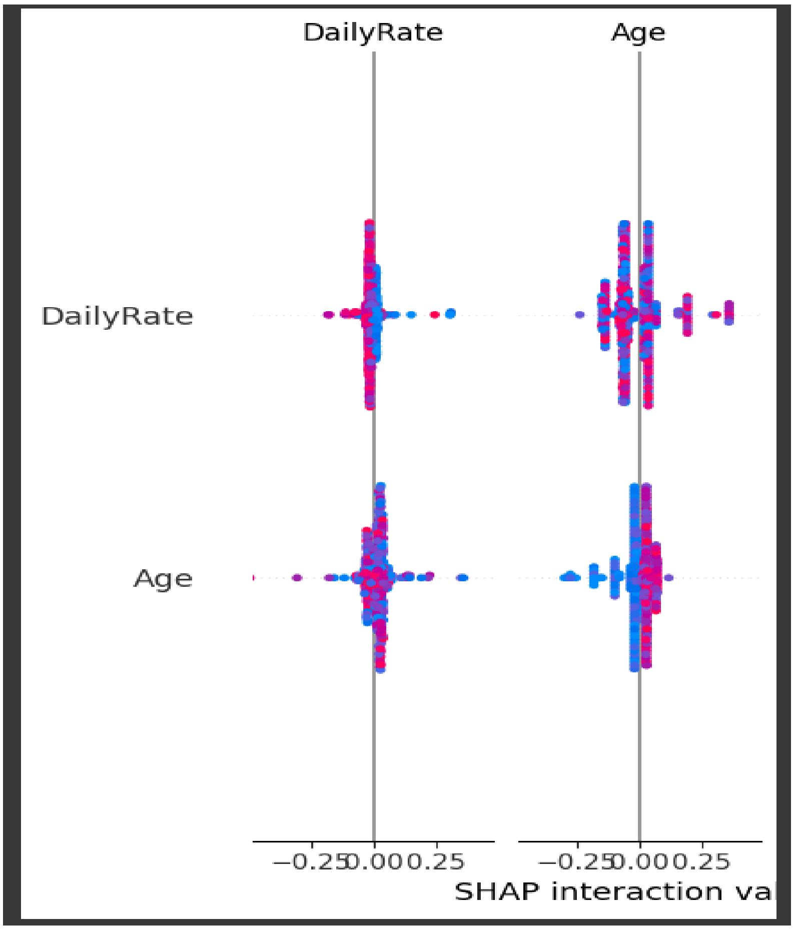
Confusion Matrix (Decision Tree):
Predicted non-attrition correctly for 238 employees but only detected 8 out of 47 actual attrition cases.



```
# Confusion Matrix

cm_dt = confusion_matrix(y_test, y_pred_dt)
plt.figure(figsize=(4,3))
sns.heatmap(cm_dt, annot=True, fmt='d', cmap='Greens')
plt.title('Confusion Matrix - Decision Tree')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```



SHAP Value Interpretation:
SHAP interaction plots showed that lower DailyRate and younger Age slightly increase the probability of attrition. The SHAP analysis provided transparency in understanding how individual features influenced model predictions.

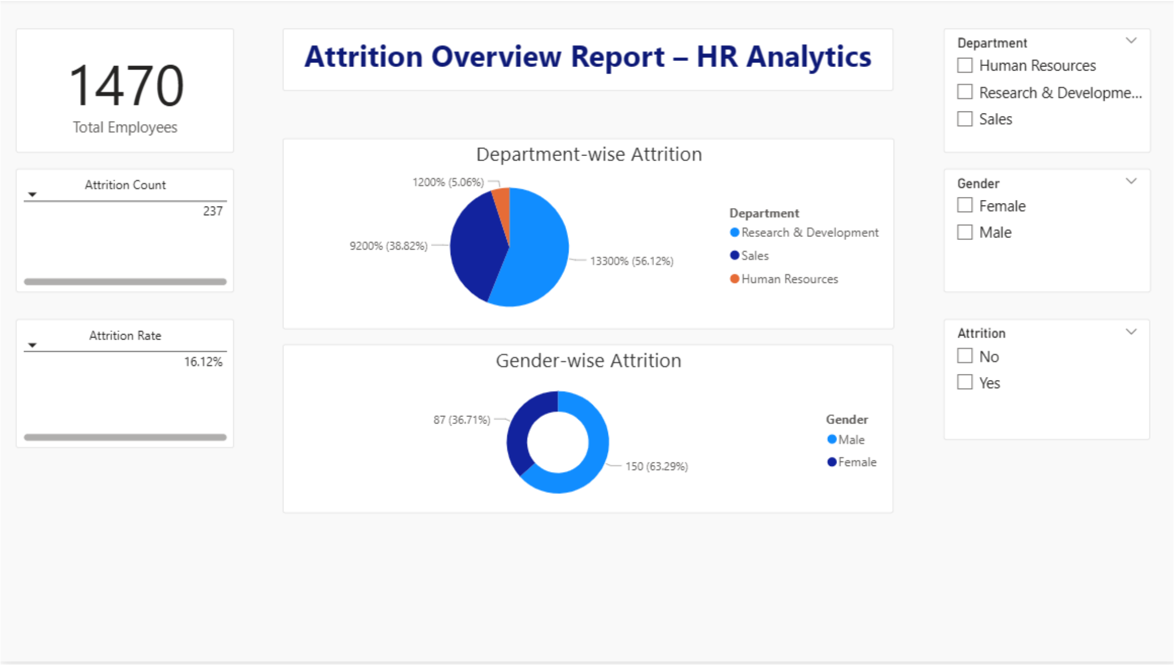


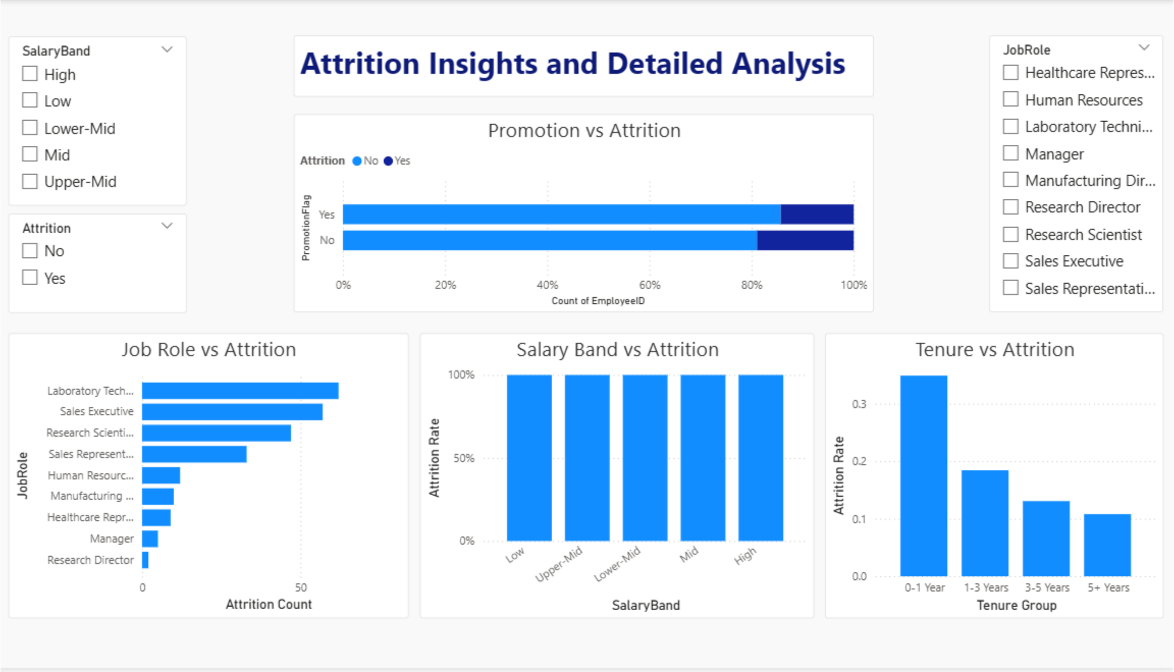
4. Conclusion

The HR Analytics project successfully identified critical factors influencing employee attrition. Logistic Regression performed better overall in prediction accuracy and interpretability compared to the Decision Tree model. However, both models indicated challenges with class imbalance, as most employees did not leave the organization.

Machine learning combined with SHAP explainability allowed the HR team to not only predict attrition but also understand the underlying reasons behind it.

PowerBI DashBoard:





5. Recommendations & Attrition Prevention Strategies

Based on the insights from the models and data analysis, the following measures are recommended to reduce attrition:

- Implement periodic salary reviews, especially for employees with low DailyRates.
- Offer career growth opportunities and transparent promotion policies.
- Improve work-life balance through flexible work arrangements.
- Identify and mentor younger employees or those in high-stress roles.
- Conduct regular employee satisfaction surveys to capture early warning signs.

This data-driven approach provides HR departments with a strategic advantage in improving employee retention and organizational stability.