Employee Attrition Prediction Project Report

1. Introduction

Employee attrition is a significant challenge for many organizations as it leads to increased recruitment costs, loss of institutional knowledge, and disruption of business operations. This project aims to develop a classification model to predict whether an employee will leave a company based on HR data, enabling HR teams to take proactive retention measures.

2. Dataset Description

The IBM HR Analytics Employee Attrition dataset contains information about employees, including demographics, job role, performance, and satisfaction levels. The dataset includes features such as Age, Gender, Job Role, Monthly Income, Years at Company, and Attrition status (Yes/No).

3. Exploratory Data Analysis (EDA)

- Attrition distribution: Approximately 16% of employees left, indicating an imbalanced dataset.
- Key factors identified:
  • Job Satisfaction and Work-Life Balance scores strongly correlate with attrition.
  • Employees with longer commute distances and higher overtime hours have higher attrition rates.
  • Low Monthly Income and fewer Years at Company are associated with leaving.
- Data imbalance: The minority class (attrition = Yes) was underrepresented, requiring careful handling.

4. Data Preprocessing

- Converted categorical variables using one-hot encoding.
- Handled missing values (if any).
- Split data into training and testing sets (80%-20%).
- Applied Synthetic Minority Oversampling Technique (SMOTE) to balance classes in the training set.

5. Model Training and Evaluation

We trained a Random Forest Classifier to predict employee attrition. The model performance on the test set is summarized below:

- Accuracy: 87.41%

- Precision: 66.67%

- Recall: 10.26%

- F1 Score: 17.78%

Classification Report:

Class | Precision | Recall | F1-Score | Support

------|-----------|--------|----------|---------

0 (No Attrition) | 0.88 | 0.99 | 0.93 | 255

1 (Attrition) | 0.67 | 0.10 | 0.18 | 39

- Accuracy: 0.87

- Macro Avg: Precision 0.77, Recall 0.55, F1-Score 0.55

- Weighted Avg: Precision 0.85, Recall 0.87, F1-Score 0.83

Although the overall accuracy is high, the recall for the attrition class is low, indicating the model struggles to correctly identify employees who will leave. This highlights the challenge posed by class imbalance in the dataset.

6. Model Explanation and Interpretation

We initially planned to use SHAP (SHapley Additive exPlanations) to explain model predictions. However, due to:
- The high computational cost of SHAP with the Random Forest on a relatively large and imbalanced dataset,
- And the class imbalance complicating SHAP's local explanation consistency,
we opted for LIME (Local Interpretable Model-Agnostic Explanations) instead. LIME provided interpretable, instance-level explanations for predictions and highlighted influential features affecting attrition likelihood.

7. Key Findings from LIME Explanations

- High overtime hours and low job satisfaction significantly increased attrition risk.
- Employees with fewer years at the company and longer commute distances were more likely to leave.
- Lack of work-life balance was a major driver for attrition in many cases.

8. Actionable Retention Strategies

Based on model insights:
- Reduce overtime and promote flexible work schedules to improve work-life balance.
- Increase job satisfaction through recognition programs, career development, and

employee engagement initiatives.
- Provide support for employees with long commutes, such as remote work options or transport allowances.
- Focus retention efforts on new employees (less tenure) through onboarding and mentoring programs.

9. Challenges Faced

- Imbalanced dataset required oversampling techniques (SMOTE) to improve model performance on the minority class.
- Choosing explainability method: SHAP was computationally expensive and less stable in this context, so LIME was used instead. This choice ensured practical, clear explanations for HR stakeholders.

10. Conclusion

This project successfully developed a classification model to predict employee attrition with interpretable insights. The use of LIME enabled HR teams to understand key attrition drivers at the individual level and design targeted interventions to reduce turnover and retain talent.