### HR Analytics Project Report

#### 1. Introduction

This project embarks on a comprehensive analysis of Human Resources (HR) data with the overarching goal of understanding the intricate dynamics that lead to employee attrition. The primary objectives of this study are multifaceted: first, to conduct an in-depth analysis of attrition rates across various departments within the organization; second, to investigate how different salary bands and promotion histories correlate with employee retention; and third, to develop a predictive classification model capable of forecasting employee attrition. Finally, the project aims to leverage SHAP (SHapley Additive exPlanations) analysis to demystify the model's predictions, providing actionable insights into the key factors driving attrition. The foundation of this analysis is the HR\_Analytics.csv dataset, which contains a rich collection of employee-related attributes.

#### 2. Data Loading and Preprocessing

The initial phase of the project involved loading the HR data into a pandas DataFrame. This was successfully accomplished by reading the HR\_Analytics.csv file from the /content/HR\_Analytics.csv path. The dataset, upon inspection, was found to comprise 38 distinct columns, encompassing a wide range of employee characteristics such as EmpID, Age, Attrition, Department, MonthlyIncome, YearsSinceLastPromotion, and many more, providing a robust basis for subsequent analysis.

The data preprocessing stage focused on preparing the raw data for effective model training:

* **Missing Value Assessment**: An initial check for missing values revealed 57 entries missing in the YearsWithCurrManager column. Although the original notebook's markdown noted no missing values, the output clearly indicated their presence. Importantly, the notebook did not include steps to explicitly handle these missing values (e.g., imputation or removal), which could be a consideration for future enhancements.
* **Categorical Variable Identification**: Several columns containing non-numerical data were identified as categorical variables. These included AgeGroup, BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, SalarySlab, Over18, and OverTime.
* **One-Hot Encoding**: To convert these categorical variables into a numerical format that machine learning algorithms can process, one-hot encoding was applied. This process creates new binary (0 or 1) columns for each unique category within a feature, effectively transforming qualitative data into a quantitative representation. The EmpID and Attrition columns were intentionally excluded from this encoding, as EmpID is a unique identifier and Attrition is the target variable.
* **Feature and Target Separation**: The preprocessed DataFrame was then logically divided into two main components: X representing the feature set (independent variables) and y representing the target variable (Attrition). For consistency in modeling, the Attrition variable was transformed into a numerical format, where 'Yes' (indicating attrition) was mapped to 1 and 'No' was mapped to 0.
* **Data Splitting**: To ensure that the developed model could generalize well to unseen data, the dataset was split into distinct training and testing sets. An 80/20 ratio was used, with 80% of the data allocated for training the model and the remaining 20% reserved for testing its performance. A random\_state of 42 was set during this split to guarantee reproducibility of the dataset partitioning. The training set X\_train had a shape of (1184, 55), X\_test (296, 55), y\_train (1184,), and y\_test (296,).

#### 3. Exploratory Data Analysis (EDA)

The exploratory data analysis phase was crucial for uncovering patterns and relationships within the HR dataset, providing initial insights into the drivers of attrition:

* **Attrition by Department**: Analysis of attrition rates across different departments revealed significant variations. The Sales department exhibited the highest attrition rate, standing at approximately 20.67%. The Human Resources department also showed a considerable attrition rate of about 19.05%. In contrast, the Research & Development department demonstrated the lowest attrition, at roughly 13.75%. These findings suggest that departmental culture, workload, or specific departmental challenges might play a role in employee retention.
* **Attrition by Salary Slab**: A clear inverse relationship was observed between salary levels and attrition rates. Employees falling into the "Upto 5k" salary slab showed the highest propensity to leave the company, with an attrition rate of approximately 25.1%. The "5k-10k" salary slab also experienced a notable attrition rate of about 15.3%. As salaries increased, attrition rates progressively decreased, with "10k-15k" at 11.5% and the "15k+" slab showing the lowest attrition at 6.25%. This strongly indicates that competitive compensation is a significant factor in retaining employees, particularly in lower salary brackets.
* **Impact of Promotions on Attrition**: The analysis of YearsSinceLastPromotion grouped by Attrition status provided compelling insights. On average, employees who chose to attrite had received their last promotion approximately 1.94 years prior. Conversely, employees who remained with the company had a longer average period since their last promotion, approximately 2.94 years. This suggests that a perceived lack of career progression or prolonged periods without promotion could be a contributing factor to employee turnover.

#### 4. Model Building and Evaluation (Details Forthcoming)

This section would typically detail the construction and assessment of the classification model aimed at predicting employee attrition. It would cover:

* **Model Selection**: The specific machine learning algorithm chosen for the attrition prediction (e.g., Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, etc.).
* **Model Training**: The process of training the selected model using the X\_train and y\_train datasets, including any hyperparameter tuning performed.
* **Model Evaluation**: A thorough evaluation of the model's performance using appropriate metrics on the unseen X\_test and y\_test data. Key metrics would typically include accuracy, precision, recall, F1-score, and potentially AUC-ROC curves, providing a comprehensive understanding of the model's predictive capabilities.

Please note: The provided Jupyter notebook content did not contain the code or output for the model building and evaluation steps, hence a detailed elaboration on this section is not possible at this time.

#### 5. SHAP Analysis (Details Forthcoming)

This section would elucidate the interpretability aspects of the attrition prediction model through SHAP (SHapley Additive exPlanations) analysis. SHAP values help explain individual predictions by quantifying the contribution of each feature to the prediction. This section would typically include:

* **Global Feature Importance**: Visualizations or discussions showcasing which features have the most significant overall impact on the model's attrition predictions across the entire dataset.
* **Local Explanations**: Examples demonstrating how SHAP values explain specific instances of attrition or non-attrition, highlighting which features pushed the prediction towards a particular outcome for an individual employee.
* **Insights into Feature Impact**: Detailed discussion on how specific features (e.g., MonthlyIncome, JobLevel, YearsAtCompany) influence the likelihood of attrition, providing actionable insights for HR strategies.

Please note: The provided Jupyter notebook content did not contain the code or output for the SHAP analysis, hence a detailed elaboration on this section is not possible at this time.

#### 6. Conclusion and Recommendations

Based on the thorough exploratory data analysis, several critical factors contributing to employee attrition have been identified. Notably, employees in the Sales and Human Resources departments, as well as those in lower salary brackets (especially "Upto 5k"), exhibit higher attrition rates. Furthermore, a discernible pattern suggests that employees who have experienced longer periods without a promotion are more inclined to leave the organization.

To proactively address and mitigate employee attrition, the following recommendations, synthesized from the available data analysis, are proposed:

* **Compensation and Benefits Review**: Prioritize a comprehensive review and potential adjustment of compensation and benefits, particularly for employees in the "Upto 5k" and "5k-10k" salary tiers. Ensuring competitive remuneration is crucial for retaining talent in these segments.
* **Enhanced Career Development and Promotion Frameworks**: Implement robust career development programs and transparent promotion pathways. The data underscores that consistent growth opportunities and timely promotions are vital for employee satisfaction and retention. Regular performance reviews should include discussions about career trajectory and potential advancement.
* **Department-Specific Retention Strategies**: Develop and deploy tailored retention strategies for the Sales and Human Resources departments. This could involve conducting departmental-specific surveys to identify unique pain points, implementing targeted engagement initiatives, or adjusting management practices to better support employees in these areas.

While the current analysis provides strong indicators, a more complete understanding and precise recommendations would be achievable with the inclusion of the model building, evaluation, and SHAP analysis sections, which would offer predictive power and deeper insights into feature importance.