**HR Attrition Analysis Report**

**1. Introduction**

Employee attrition is a critical concern for organizations as it affects productivity, morale, and costs. This report details a data-driven analysis using an HR attrition dataset. The goal is to explore key factors influencing employee turnover and build a predictive model that can identify employees at higher risk of leaving. Our approach includes data preprocessing, exploratory analysis, correlation studies, and the application of a logistic regression model.

**2. Data Preprocessing**

**2.1 Data Loading and Inspection**

* **Dataset Source:**  
  The dataset is provided as a CSV file (WA\_Fn-UseC\_-HR-Employee-Attrition.csv).
* **Initial Inspection:**  
  After loading the data using pandas, basic information such as the shape of the dataset and the list of columns is obtained. This helps in understanding the structure and identifying key features.

**2.2 Handling Categorical Variables**

* **Target Variable Encoding:**  
  The target variable "Attrition" (values "Yes"/"No") is encoded into a binary numeric format using a LabelEncoder, resulting in a new column Attrition\_enc (1 for "Yes", 0 for "No").
* **Encoding Other Categorical Features:**  
  Categorical features, such as "OverTime", are also encoded into numeric values to allow them to be used in modeling.

**2.3 Feature Selection**

* **Selected Features for Modeling:**  
  For the predictive model, a subset of features was chosen based on domain relevance:
  + Age
  + DistanceFromHome
  + MonthlyIncome
  + JobSatisfaction
  + YearsAtCompany
  + NumCompaniesWorked
  + OverTime

These features were selected because they are commonly linked to employee engagement and retention.

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

**3.1 Descriptive Statistics**

* **Numeric Features:**  
  The describe() function was used to generate summary statistics (mean, median, standard deviation, etc.) for numeric features. This provided insight into the distribution and scale of the data.
* **Categorical Features:**  
  Summary statistics for categorical variables (such as counts and unique values) were obtained to understand the composition of the dataset.

**3.2 Visualizations**

Several plots were generated to visualize the data:

* **Attrition Count Bar Chart:**  
  A bar chart was created to show the count of employees who stayed versus those who left, providing a quick view of class imbalance.
* **Age Distribution Histogram:**  
  A histogram illustrated the distribution of employee ages, helping to determine if certain age groups are more prone to attrition.
* **Monthly Income Box Plot:**  
  A box plot was used to compare monthly incomes between employees who left (Attrition = Yes) and those who stayed (Attrition = No). This visualization helped to reveal if income differences might be associated with attrition.

**3.3 Correlation Analysis**

* **Correlation Matrix & Heatmap:**  
  A correlation matrix was computed using all numerical variables, including the encoded attrition column (Attrition\_enc). A heatmap visualized the strength and direction of correlations between variables.
* **Key Findings from Correlation:**
  + **OverTime:**  
    A strong positive correlation (among the highest observed) indicates that working overtime is associated with higher attrition.
  + **JobSatisfaction and Related Factors:**  
    Negative correlations with attrition suggest that lower job satisfaction contributes to higher turnover.
  + **Experience and Tenure:**  
    Variables such as TotalWorkingYears, YearsAtCompany, and Age showed negative correlations with attrition, implying that more experienced or longer-tenured employees tend to be more stable.

**4. Modeling Techniques**

**4.1 Predictive Modeling Approach**

* **Objective:**  
  Build a model to predict which employees are at higher risk of attrition.
* **Algorithm Chosen:**  
  Logistic Regression was selected due to its interpretability and suitability for binary classification tasks.

**4.2 Data Splitting and Model Training**

* **Train-Test Split:**  
  The dataset was split into training and testing sets (70% for training and 30% for testing) to evaluate the model's performance on unseen data.
* **Model Training:**  
  The logistic regression model was trained on the selected features. The training process adjusted the model coefficients to best fit the training data.

**4.3 Model Evaluation**

* **Accuracy:**  
  The model achieved an overall accuracy of approximately 84.6%.
* **Classification Report:**
  + **Majority Class (No Attrition):**  
    Precision of 0.86 and recall of 0.97 indicate that the model performs well for predicting employees who stay.
  + **Minority Class (Attrition):**  
    The model had a low precision (0.23) and recall (0.05) for predicting attrition, highlighting a significant challenge due to class imbalance.
* **Feature Coefficients:**  
  The coefficients indicate the influence of each feature on the probability of attrition:
  + **OverTime (1.54):** Strong positive influence, meaning employees working overtime are much more likely to leave.
  + **JobSatisfaction (-0.28):** Negative coefficient suggests that higher satisfaction reduces the likelihood of leaving.
  + **Other Features:**  
    Variables like Age, YearsAtCompany, and DistanceFromHome also contribute but with smaller magnitudes.

**4.4 Challenges and Considerations**

* **Class Imbalance:**  
  The dataset has a significant imbalance (more employees staying than leaving), which results in the model being less effective at predicting the minority class (attrition).
* **Future Improvements:**  
  To better capture the minority class, techniques such as oversampling (e.g., SMOTE), undersampling, or adjusting class weights could be applied. Additionally, exploring more complex models (like Random Forests or Gradient Boosting) might improve predictive performance.

**5. Results and Discussion**

**5.1 Model Performance Summary**

* **Accuracy:** 84.6% overall.
* **Precision & Recall:**
  + For the majority class (non-attrition), the model performs well.
  + For the attrition class, very low recall (5%) and low precision (23%) indicate that many at-risk employees are not being correctly identified.

**5.2 Interpretation of Feature Impact**

* **OverTime:**  
  The high coefficient reinforces that overtime work is a major risk factor for attrition.
* **JobSatisfaction:**  
  Improvements in job satisfaction could potentially reduce turnover.
* **Experience-Related Factors:**  
  Increasing tenure and experience are associated with reduced attrition risk.

**5.3 Actionable Insights**

* **Review Overtime Policies:**  
  Since overtime is a strong predictor of attrition, organizations should consider strategies to balance workloads.
* **Enhance Employee Engagement:**  
  Addressing factors that drive job dissatisfaction—such as through career development programs and regular feedback—can help retain employees.
* **Address Class Imbalance:**  
  Future models should focus on techniques to better predict the minority class, ensuring at-risk employees are identified for early intervention.

**6. Conclusion**

This analysis provides a comprehensive view of factors influencing employee attrition and highlights the challenges of predicting turnover in an imbalanced dataset. While the logistic regression model achieves reasonable overall accuracy, its low performance on the attrition class calls for further refinement through advanced sampling techniques and possibly more complex models.

By addressing key drivers like overtime and job satisfaction, HR can develop targeted retention strategies. Continuous monitoring, enhanced feature engineering, and model updates will be essential for improving predictive performance and ultimately reducing employee turnover.