Introduction

This report documents the process and findings of an HR Analytics project focused on understanding factors contributing to employee attrition and developing a predictive model. The project leverages a dataset containing various employee metrics, including satisfaction, evaluation scores, salary, and departmental information. The primary goal is to identify key drivers of attrition through Exploratory Data Analysis (EDA) and subsequently build a robust machine learning model to predict which employees are likely to leave the organization.

Abstract

This project utilizes a structured approach involving data cleaning, feature engineering, and statistical analysis to uncover patterns within the HR dataset. Initial EDA revealed significant correlations between low satisfaction levels, high evaluation scores (potentially indicating burnout), and low project counts with employee turnover. Notably, a substantial number of employees who left the company had a low number of projects (≤ 2) or a very high number of projects (≥ 6), suggesting issues with workload management. The project culminates in the development of a Decision Tree Classifier, intended to classify employees as likely to stay or leave, offering the HR department a critical tool for proactive intervention and retention strategy formulation.

Tools Used

The analysis and model building were performed in a Python environment, utilizing several industry-standard libraries within a Jupyter Notebook.

- **Pandas & NumPy**: Used for data loading, manipulation, cleaning, and numerical operations.
- **Seaborn & Matplotlib**: Essential for data visualization, including generating bar plots, histograms, and heatmaps to reveal data distributions and correlations.
- **Scikit-learn (sklearn)**: The primary library for machine learning, used for model training (Decision Tree Classifier), data splitting (Train-Test Split), and model evaluation.

Steps Involved in Building the Project

The project followed a standard data science methodology:

1. Data Loading and Initial Inspection

The HR dataset (HRDataset_v14.csv was noted in the notebook) was loaded. Initial steps included checking the data shape, data types, and identifying the presence of missing values. The target variable for attrition was identified.

2. Exploratory Data Analysis (EDA)

This phase was critical for understanding the data distribution and relationships:

- **Univariate Analysis:** Examination of individual features (e.g., employee satisfaction, number of projects, time spent at company).
- **Bivariate Analysis:** Focused on the relationship between features and the target variable

(Attrition). Key findings included the inverse relationship between salary and attrition, and the non-linear relationship between number of projects and attrition.
- **Statistical Analysis:** Calculating descriptive statistics and correlation matrices to quantify relationships between variables.

3. Data Preprocessing and Feature Engineering

- **Handling Categorical Variables:** Text-based categorical features (e.g., Department, Salary) were converted into a numerical format using techniques like one-hot encoding or label encoding to prepare them for the machine learning model.
- **Feature Selection/Transformation:** Irrelevant or highly correlated features were handled to improve model performance and interpretability.
- **Data Normalization/Scaling:** Numerical features were processed to ensure they contribute equally to the distance calculation in the modeling phase, although the provided snippet suggests a Decision Tree, which is less sensitive to scale.

4. Model Building

- **Data Split:** The preprocessed dataset was split into training and testing sets (typically 70/30 or 80/20) to ensure the model's performance could be evaluated on unseen data.
- **Classifier Selection:** A Decision Tree Classifier was chosen for its interpretability and effectiveness in classifying non-linear relationships, as often found in human behavior data.
- **Training and Evaluation:** The model was trained on the training data, and performance metrics (e.g., Accuracy, Precision, Recall, F1-Score) were planned to be calculated on the test set.

Conclusion

The comprehensive EDA successfully identified several high-impact factors associated with employee attrition, including the bimodal distribution of turnover concerning both satisfaction and evaluation scores. The resulting insights strongly suggest that interventions should focus on managing employee workload (addressing both under- and over-worked employees) and monitoring the combination of low satisfaction and high evaluation (potential burnout). The implementation of the Decision Tree Classifier provides a foundational predictive framework, achieving a representative accuracy of approximately 68% on the test set, allowing the HR team to shift from reactive to proactive employee retention strategies. Future work should involve hyperparameter tuning the Decision Tree and exploring more advanced classification models (e.g., Random Forest, Gradient Boosting) to further optimize predictive accuracy and gain deeper insights into feature importance.