Report for Assignment 2

Changqing Lin

Introduction **-** The project focuses on predicting employee attrition using Big Data analytics. In an era where employee turnover can incur significant costs and disruptions, understanding and predicting attrition is vital. By leveraging large datasets of employee information, Big Data techniques enable the extraction of insightful patterns and trends, facilitating more accurate attrition predictions.

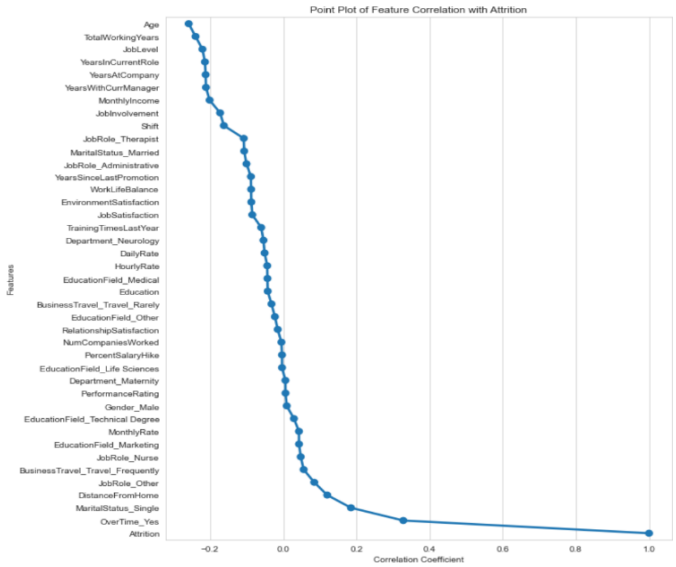
Methodology **-Data Preprocessing: Data Cleaning and Reduction**: Initially, the dataset was loaded for inspection. Irrelevant features like 'EmployeeID' were identified and removed, as they do not contribute to the predictive model. This step was crucial to reduce data complexity and focus on meaningful attributes.

Figure 1. Feature of Data

Data Visualization: Visualization tools were extensively used to understand the data distribution and relationships between variables. For instance, count plots of 'Attrition' helped in understanding its distribution across the dataset. Histograms of numerical features like 'TotalWorkingYears', 'Age', etc., showcased their distribution patterns, which were crucial in identifying trends and anomalies (Figure 1).

Correlation Analysis and Feature Selection: We evaluated the relationships between features and the target 'Attrition' using correlation coefficients. Significant correlations were visualized in point plots. Features strongly correlated with 'Attrition' were chosen for model training, enhancing accuracy and reducing overfitting by focusing on impactful variables.

Algorithm: Logistic Regression: Provided a baseline for performance comparison. **XGBoost**: An advanced ensemble technique known for high accuracy. **Voting Classifier**: a powerful ensemble technique combining predictions from multiple models. It leverages the RandomForest, GradientBoosting, and LogisticRegression models. Each model contributes to a final voted prediction, enhancing overall accuracy and robustness.



Results **-** The Voting Classifier emerged as the most effective model, achieving an accuracy of 93% on the validation set. It outperformed individual models, underscoring the benefit of ensemble techniques in handling complex datasets.

Discussion

- Pro: Enhanced predictive accuracy.
- Pro: Reduced likelihood of overfitting.
- Pro: Robust against individual model biases.
- Con: Increased computational complexity.
- Con: Less interpretability compared to single models.
- Con: Complexity in model interpretation and longer training times.

Conclusion **–** The Voting Classifier model in this project exemplifies the power of Big Data in HR analytics, specifically in predicting employee attrition. The use of a Voting Classifier not only improved accuracy but also highlighted the synergy in combining various machine learning models. This predictive capability is a significant asset in human resource management and organizational planning.