Report on IBM HR Attrition Dataset

Introduction:

The IBM HR Attrition Dataset is a valuable source of information for understanding the factors that influence employee attrition in the corporate world. In this report, we analyse the data to identify patterns and trends that could inform HR practices and help organizations reduce employee turnover rates.

Data Description:

The dataset consists of employee records of a certain company. It contains 35 columns and 1470 rows. The columns are Age, Attrition, business travel, daily rate, Department, distance from home, Education, education field, employee count, employee number, environment satisfaction, Gender, hourly rate, job involvement, job level, job role, job satisfaction, marital status, monthly income, monthly rate, number of companies worked, Over18, overtime, percentage salary hike, performance rating, relationship satisfaction, standard hours, stockoptionlevel, totalworkingyears, trainingtimeslastyear, worklifebalance, yearsatcompany, yearsincurrentrole, yearssincelastpromotion, and yearswithcurrmanager.

The "Attrition" column is the target variable, and it indicates whether an employee has left the company or not. The dataset contains both numerical and categorical variables. The numerical variables include Age, daily rate, distance from home, Education, environment satisfaction, hourly rate, job involvement, job level, job satisfaction, monthly income, monthly rate, number of companies worked, percentage salary hike, performance rating, relationship satisfaction, standard hours, stockoptionlevel, totalworkingyears, trainingtimeslastyear, worklifebalance, yearsatcompany, yearsincurrentrole, yearssincelastpromotion, and yearswithcurrmanager. The categorical variables include business travel, Department, education field, Gender, job role, marital status, Over18, and overtime.

The dataset has no missing values and duplicates. The data has been pre-processed, including encoding categorical variables, standardizing the

numerical variables, and removing unnecessary columns. This dataset will be used for the prediction of employee attrition using machine learning models.

Exploratory Data Analysis:

The exploratory data analysis (EDA) reveals that the dataset is well-balanced, with 1233 employees who have not left the company (83.88%) and 237 employees who have left (16.12%). The analysis also shows that most of the employees are between the ages of 25 and 45, have an education level of Bachelor's Degree, and have been with the company for less than 10 years.

Furthermore, the EDA identifies several factors that influence employee attrition, such as low job satisfaction, low monthly income, and high overtime hours. Additionally, the EDA highlights the importance of job-related factors, such as job level, job role, and department, in predicting employee attrition.

Machine Learning:

We trained a logistic regression model to predict employee attrition using the IBM HR Attrition Dataset. The model achieved an accuracy of 80%, which is a good result. The confusion matrix shows that the model correctly predicted 144 employees who did not leave the company and 4 employees who left the company. However, the model incorrectly predicted that 29 employees who left the company did not leave and 0 employees who did not leave the company left.

Deep Learning:

We also trained a deep learning model to predict employee attrition using the IBM HR Attrition Dataset. The model achieved an accuracy of 78%, which is a reasonable result. The scores for the model evaluation indicate that the model has a good precision rate, but a low recall rate, which means that the model is better at identifying employees who will stay rather than those who will leave the company.

Conclusion:

The IBM HR Attrition Dataset provides valuable insights into the factors that influence employee attrition in organizations. Our analysis shows that job-related factors, such as job level, job role, and department, are important predictors of employee attrition. Additionally, the EDA highlights the

importance of addressing job satisfaction, monthly income, and overtime hours to reduce employee turnover rates.

The machine learning and deep learning models we trained provide a good foundation for predicting employee attrition. The logistic regression model achieved a good accuracy rate, while the deep learning model achieved a reasonable accuracy rate with good precision but low recall. However, it is important to note that these models were trained on a limited dataset, and additional data may be needed to improve their accuracy.

Recommendations: Based on our analysis, we recommend that organizations focus on addressing job-related factors, job satisfaction, monthly income, and overtime hours to reduce employee attrition rates. Furthermore, we recommend that organizations continue to collect data on employee turnover rates and use machine learning models to identify trends and patterns in the data that can inform HR practices.

In conclusion, the IBM HR Attrition Dataset provides valuable insights into the factors that influence employee attrition, and organizations can use this information to reduce employee turnover rates and improve employee retention.

Researcher: sabune Radhika

Qualification: Master's in Statistics

Thank you