Case Study: Predicting Employee Performance in a Company

Objective

As a data analyst for a Human Resources (HR) department, your task is to build a linear regression model to predict employee performance based on various factors. This model will help HR managers identify high-performing employees, improve talent management strategies, and provide insights into factors that influence employee success.

Dataset

The dataset could include the following fields:

- **Performance Score**: The employee's annual performance rating or score (target variable)
- **Age**: The age of the employee
- **Years of Experience**: Total years of work experience
- **Education Level**: The highest level of education attained (e.g., Bachelor's, Master's, PhD)
- **Hours Worked Per Week**: Average number of hours the employee works each week
- **Job Role**: The role or job title of the employee (e.g., Software Developer, Marketing Specialist)
- **Team Size**: The number of people on the employee's team
- **Training Hours**: The number of hours the employee has spent on training during the year
- **Previous Performance**: Performance score from the previous year
- **Absenteeism**: The number of days an employee was absent from work
- **Salary**: The salary of the employee
- **Job Satisfaction**: A rating of job satisfaction (e.g., from 1 to 5)
- **Location**: The location where the employee works (can be encoded as categorical values)

You can find similar datasets on HR analytics platforms like Kaggle, or generate a synthetic dataset.

Steps to Solve the Case Study

1. **Data Exploration and Preprocessing**
    - Check for missing values and handle them appropriately through imputation or removal.
    - Convert categorical variables like Job Role and Location to numerical format using encoding (e.g., one-hot encoding).
    - Ensure that the data types of each feature are appropriate for the analysis (e.g., numerical features like Salary, Years of Experience).
2. **Exploratory Data Analysis (EDA)**

- Use scatter plots to visualize the relationship between Performance Score and other numerical features such as Years of Experience, Training Hours, and Hours Worked Per Week.
- Analyze correlations to understand which factors have the highest relationships with employee performance.
- Visualize distributions of job satisfaction, absenteeism, and other features to assess their potential influence.

3. **Feature Engineering**
    - Create new features such as:
        - **Performance Trend**: Difference between current and previous performance scores.
        - **Work-Life Balance**: Calculate from Hours Worked Per Week and Absenteeism.
        - **Experience-to-Training Ratio**: Ratio of Years of Experience to Training Hours, which may indicate how experience and learning impact performance.
4. **Split the Data**
    - Divide the dataset into training and testing sets (typically an 80-20 or 70-30 split) to validate the model's performance on unseen data.
5. **Build the Linear Regression Model**
    - Fit a linear regression model with Performance Score as the target variable.
    - Evaluate the significance of each feature using p-values and examine confidence intervals.
6. **Evaluate the Model**
    - Use performance metrics such as **R-squared**, **Mean Absolute Error (MAE)**, and **Mean Squared Error (MSE)** to assess how well the model predicts employee performance.
    - Check residuals for randomness to ensure that the model assumptions are met.
7. **Interpret Results and Insights**
    - Identify which features are the most influential in predicting performance (e.g., training hours, previous performance, job satisfaction).
    - Explore how factors like age, experience, or absenteeism might negatively or positively affect performance predictions.
8. **Recommendations for HR Strategy**
    - Based on the analysis, provide insights and actionable recommendations for improving employee performance, such as:
        - Offering additional training to employees with low training hours.
        - Targeting employees with poor past performance scores for coaching.

        - Understanding the relationship between absenteeism and performance to implement attendance policies.

9. **Communicate Results**
    - Prepare a report or dashboard with visualizations showing key insights, such as performance distribution across departments or the impact of experience and training on performance.
    - Provide strategic recommendations based on model findings for HR to take actionable steps to improve performance management.