**Employee Performance Analysis**

**INX Future Inc.**

**PROJECT SUMMARY:**

**WE NEED TO PREDICT THE PERFOMANCE RATING OF EMPLOYEE**

**INX Future Inc Employee Performance – Project**

**The following insights are expected from this project.**

1. Department wise performances
2. Top 3 Important Factors effecting employee performance
3. A trained model which can predict the employee performance based on factors as inputs. This will be used to hire employees
4. Recommendations to improve the employee performance based on insights from analysis.

**Business Case:**

To analyses the current employee data and find the core underlying causes of this performance issues. Expects the findings of this project will help to take right course of actions. Also expects a clear indicators of non-performing employees, so that any penalization of non-performing employee, if required, may not significantly affect other employee morals.

The given Employee dataset consist of 1200 rows. The features present in the data are 28 columns. The shape of the dataset is 1200x28. The 28 features are classified into quantitative and qualitative where 19 features are quantitative (11 columns consists numeric data & 8 columns consists ordinal data) and 8 features are qualitative. EmpNumber consist alphanumerical data (distinct values) which doesn't play a role as a relevant feature for performance rating.

# Importing necessary libraries

1. Pandas
2. Numpy
3. Seaborn
4. Matplotlib
5. Warnings

**Pandas:** Pandas is a powerful data manipulation and analysis library. It provides data structures (like Data Frames and Series) to efficiently store and work with structured data. Pandas is widely used for tasks like data cleaning, transformation, and data exploration.

**Numpy:** NumPy stands for Numerical Python. It provides support for working with arrays and matrices, as well as a large collection of mathematical functions to operate on these data structures. It's a fundamental library for numerical computations in Python.

**Seaborn:** Seaborn is a statistical data visualization library based on Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is particularly useful for creating complex visualizations and is often used in conjunction with Pandas.

**Matplotlib:** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It's a fundamental tool for creating plots, graphs, and other visual representations of data. Matplotlib is often used in combination with other libraries like NumPy and Pandas.

**Warnings:** The warnings module in Python provides a way to deal with warnings in your code. Warnings are messages that indicate something unexpected or potentially problematic is happening, but doesn't necessarily prevent the program from running. It's important to be aware of warnings, as they can signal potential issues in your code.

These libraries are essential for anyone working with data in Python. They provide the tools needed to manipulate, analyze, and visualize data effectively. Each one serves a specific purpose and can be used in conjunction with others to perform complex data science tasks.

**Basic Check & Statistical Measures**

There is no constant column is present in Numerical as well as categorical data.

# Numerical columns

1. Age
2. DistanceFromHome
3. EmpEducationLevel
4. EmpEnvironmentSatisfaction
5. EmpHourlyRate
6. EmpJobInvolvement
7. EmpJobLevel
8. EmpJobSatisfaction
9. NumCompaniesWorked
10. EmpLastSalaryHikePercent
11. EmpRelationshipSatisfaction
12. TotalWorkExperienceInYears
13. TrainingTimesLastYear
14. EmpWorkLifeBalance
15. ExperienceYearsAtThisCompany
16. ExperienceYearsInCurrentRole
17. YearsSinceLastPromotion
18. YearsWithCurrManager
19. PerformanceRating

# Categorical columns

1. EmpNumber
2. Gender
3. EducationBackground
4. MaritalStatus
5. EmpDepartment
6. EmpJobRole
7. BusinessTravelFrequency
8. OverTime
9. Attrition

# Univariate, Bivariate & Multivariate Analysis

Hist plot & Count plot is used in **Univariate** and also analysis with **sweetviz.**

Bar plot & KDE plot is used in **Bivariate** analysis.

Pair plot is used in **Multivariate** analysis.

Used Label encoding to convert categorical columns to numerical one.

# checking outliers using boxplot

Some columns contains outliers. They can affect the accuracy of the models and lead to incorrect predictions.s So we are going to find these outliers using Winsorization method and remove it. The name of columns which contain outliers are given below.

NumCompaniesWorked , TotalWorkExperienceInYears, TrainingTimesLastYear,

ExperienceYearsAtThisCompany, ExperienceYearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, PerformanceRating .

**Winsorization steps:**

name = "Name of coulumn"

df = data

# Create a figure with two subplots

fig, axes = plt.subplots(1, 2, figsize=(10, 5))

# Before winsorization

sns.boxplot(data=df[name], ax=axes[0],orient='h') # Use the specific column

axes[0].set\_title("{} Before Winsorization".format(name))

# Winsorization

winsorized\_data = winsorize(df[name], (0, 0.06)) # Use the specific column

# After winsorization

sns.boxplot(winsorized\_data, ax=axes[1],orient='h')

axes[1].set\_title("{} After Winsorization".format(name))

# Replace data in dataset

data[name] = winsorized\_data

# Machine learning Model Creation & Evaluation

1. Define Dependent and Independent Features:
2. **Balancing the data**: The data is imbalance, so we need to balance the data with the help of SMOTE

**SMOTE**: SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesis new minority instances between existing minority instances.

3.Splitting Training And Testing Data: 80% data use for training & 20% data used for testing.

**The trained model is created using the machine learning algorithm as follows with the accuracy score,**

* AdaBoostClassifier = 85.41
* RandomForestClassifier=92.50
* LogisticRegression=71.25
* GradientBoostingClassifier=90.41
* KNeighborsClassifier=48.33
* SVC=71.25

Random Forest Classifier & Gradient Boosting Classifier gives the best accuracy compared to other models.

# Department wise performance

**Sales:** The Performance rating level is 2.8 in the sales department. Male employees performance is more compared to female employees in sales.

**Human Resources:** The most of the employees were under the level 3 performance. The female employees in HR department are performing well compared to male employees. Performance of employees between the age 50- 60 performance were very low in this department.

**Development:** This department has highest level of employee performance. Employees in all age wise and gender wise performance is equal.

**Data Science:** This department has the second highest performance rate. Male employees are working more efficiently in this department. The overall performance is good compared to all departments.

**Research & Development:** The overall performance of R&D department is good. Employees in all ages working very efficiently.

**Finance:** The finance department performance is decreasing when age increases. The male employees are doing good. The experience factor is inversely relating to the performance level.

# Top 3 factors effecting employee performance

1. Employment Environment Satisfaction
2. Employee Salary Hike Percentage
3. Employee Work Life Balance

# A trained model which can predict employee model

1. Random Forest classifier: 92.50% accuracy
2. Gradient Booster: 90.40% accuracy

# Recommendations to improve the employee performance

1. The overall employee performance can be achieved by employee environment satisfaction.

The company needs to focus more on the employee environment satisfaction.

1. The salary hike will give the boost to the employees to perform well.
2. Promote the employee every 6th month.
3. Salary hike, Promotion and w
4. Improve Employee's work-life balance this affects the performance rating.
5. While recruiting for HR, consider the female candidates where they perform well compared to male.
6. The development and sales department are having an overall higher performance comparing to rest of the departments.