Employee Performance Analysis

INX Future Inc.

Project Summary:

The aim of the project is to identify the performance of the employee depending on the features given. The given dataset contains of 1200 rows/records and 27 columns, in that, we have 19 columns of numerical data ie., quantitative and 8 columns are of categorical data ie., qualitative.

The insights of the project are as follows;

- Department wise performance.
- Top 3 factors affecting the employee performance.
- A trained model which can predict the employee performance based on factors as inputs. This will be used to hire employees
- Recommendations to improve the employee performance based on insights from analysis.

The analysis of the project has gone through the stage of distribution analysis, correlation analysis and analysis of each department to satisfy the project goal. The machine learning model which is used in this project is **RandomForest Classifier** which predicts highest accuracy - 92%. The other machine learning algorithms which give high accuracy are **XGBoost Classifier** - 91% and **DecisionTree Classifier** – 90%. The overall project was performed and achieved the goals by using the machine learning model and visualization techniques.

Requirement:

The dataset "INX Future Inc" for this project is given by IABAC. This project was done in Jupyter Notebook using Python.

Analysis:

Our analysis depicts the relationship between the independent variables and dependent variable. The features present in this dataset are divided into

numerical and categorical data. The categorical data were converted into numDataerical data by using LabelEncoder.

In some features, outliers were detected and those outliers were replaced with median. Analysis was done with outliers and without outliers. There were no significant changes in the accuracy when compared.

Data Cleaning:

Data cleaning is considered as one of the main parts in Data Science Project. Before training the model, our data should be clean and structured. Our dataset has no missing values or duplicate values.

Data Visualization:

Data visualization is performed in both the ways ie., by distributing the data and by using correlation heatmap to visualize the analysis.

i) Distribution Plots:

We can explore the data by distributing the features with one another. Data Distribution was performed by using Seaborn and Matplotlib libraries widely for both numerical and categorical data.

ii) Correlation Heatmap:

The Correlation Heatmap is used to find the correlation between the features. By plotting the correlation matrix, we can have a good overview of how the features are related to each other. It also helped in finding the features that are significantly correlated with the target variable.

Conclusion:

The insights that were asked from this analysis are;

1) Department Wise Performances:

In department wise performance analysis, we have to analyse the data from each department present in the category. The data frame has to be separated or sliced according to department wise. In EmpDepartment feature there were six departments namely **Sales,**

Human Resources, Development, Data science, Reasearch & Development and **Finance.** The performance analysis by the department is as follows:

- **Sales:** The Performance rating level 2 is more in the sales department.
- **Human Resources:** The majority of the employees lying under the level 2 performance.
- **Development:** The highest average of level 3 performance is in Development department. The overall performance is higher compared to all departments.
- **Data Science:** Data science is the only department where a smaller number of level 2 performers present.
- **Research & Development:** The majority of the employees lying under the level 2 performance.
- **Finance:** Here also the majority of the employees lying under the level 2 performance. Performance rating of level 3 & 4 are very less in numbers.

2) Top 3 factors affecting employee performance:

From Correlation Heatmap we get the important features that are highly related to the target variable ie., PerformanceRating. The top three important features affecting the performance rating are ordered with their importance level as follows,

1. EmpEnvironmentSatisfaction
2. EmpLastSalaryHikePercent
3. YearsSinceLastPromotion

EmpEnvironmentSatisfaction and EmpLastSalaryHikePercent are positively correlated to the PerformanceRating whereas YearsSinceLastPromotion is negatively correlated to the PerformanceRating.

3) A Trained Model which can predict the employee performance:

The trained models are created with the Machine Learning Algorithms with the accuracy score as follows;

- Logistic Regression: 83%
- XGBoost Classifier: 91%
- DecisionTree Classifier: 90%
- RandomForest Classifier: 92%
- Naïve Bayes Bernoulli: 78%

- Support Vector Machine: 74%
- KNN Algorithm: 85%

4) Recommendations to improve the employee performance:
   - Employee Environment Satisfaction is positively correlated to the PerformanceRating which means when company provides a better environment for employees it increases the performance of the employee.

   - Employee Last Salary Hike Percent is also positively correlated to PerformanceRating. So, if company increase salary from time to time, that will increase the performance of the employee.

   - Years Since Last Promotion is negatively correlated to the PerformanceRating. So, if employee is not promoted to next level for many years, that will decrease the performance of the employee. So, company should consider promoting employees to the next level when they are eligible and should also encourage them in achieving it.

   - Employee Work Life Balance also has positive correlation with the target variable. When it increases, the performance rating of employee increases. So, company has to provide supportive and healthy work environment, which will enable them to have proper balance between job and personal responsibilities.